

# THIS WEEK

## EDITORIALS

**OCEANS** Some mysteries of the deep can be solved — if there is a will **p.244**

**WORLD VIEW** It is time to stop the silliness of science citation counts **p.245**



**ORIGINS** Maize took the high road from Mexico to the United States **p.247**

## Science and satire

*The terrorist attacks in Paris were an assault on the fundamental values of free and democratic societies. Researchers, and humorists, must combat obscurantism everywhere.*

In 1766, Jean-François Lefebvre became the last person in France to be executed for blasphemy; his list of ‘crimes’ included a refusal to remove his hat as a religious procession passed by. Eighteenth-century France was one of the first nations to push back against the tyranny of religious authority that stifled free thought across Europe at the time — and continues to do so in other places. That proud legacy — and the part that both science and satire played in promoting the contrasting values of the Enlightenment — is worth reflecting on as millions across the world struggle to make sense of the horrific events in Paris last week.

The terrorists who murdered 17 people, including 8 staff members of the French weekly satirical magazine *Charlie Hebdo*, falsely claimed to act in the name of Islam. On the contrary, the perpetrators represent a fanaticism that would stifle freedoms and science in the Arab and Islamic world, and beyond. The means used in the eighteenth century remain among the best options to combat this warped world view today. Free scientific thinking and satire — both religious and political — are crucial in challenging and undermining dogma and authoritarianism.

It is no coincidence that the eighteenth-century French writer, intellectual and activist Voltaire, a leading Enlightenment figure, was both an outspoken and irreverent satirist of religion and a leading proponent of the natural sciences as the successor to religion and philosophical reasoning as the main route to knowledge. The Enlightenment culminated in the French Revolution, and the resulting 1789 Declaration of the Rights of Man and of the Citizen, which explicitly protected free speech, including the right to freely criticize religious views.

Philippe Val, a former editor of *Charlie Hebdo*, has compared satire to science, saying that both are methods of pursuing ‘truth’ in the face of dogma. *Charlie Hebdo* is no Voltaire, but Val has a point. Satire, wit and mockery remain surprisingly effective ways to voice dissent, and to highlight the absurdities, hypocrisy, injustice and oppression of authoritarian regimes and religious obscurantism.

### FREEDOM OF THOUGHT

Satire is widely credited with helping to undermine the authority of regimes during the revolutions of the Arab Spring, starting in 2011. It continues to be a powerful means to air grievances and call for further freedoms not only in Tunisia, where the revolution has prompted a relatively successful transition to democracy, but also in other Arab countries where uprisings have so far ended in chaos.

Although satire may survive, in the Arab world the development of science has long been hampered. The landmark *Arab Human Development Report 2003*, written by Arab scholars and sponsored by the United Nations Development Programme, identified authoritarianism, and the resulting lack of open democratic societies that uphold freedom of speech, as the main obstacle. Greater freedoms, and science itself, it argued, work together to build free and creative societies.

The Arab spring raised hope among scientists in the region that

greater freedoms and democracy would usher in societies based on science and other forms of knowledge. The same fanaticism that brought bloodshed to the streets of Paris jeopardizes those hopes, and must be resisted.

The Paris killings are also a reminder that the hard-won freedoms that most readers of *Nature* enjoy must not be taken for granted, but must be continually defended. In many countries, it is taken as given that one is free to express opinions, criticize government policy and research any subject without worrying about intimidation, retaliation or imprisonment, much less being shot in the head.

**“Freedoms must not be taken for granted, but must be continually defended.”**

Those rights are increasingly being eroded, however — ironically, often in political over-reactions to terrorism — including through anti-terrorism laws that roll back civil liberties, increasingly invasive surveillance states and government oppression of legitimate dissent such as the Occupy movement. Politics and

other forces can sometimes lead to censorship of lines of enquiry — such as gun research in the United States — as well as more insidious forms of influence on areas of research.

Scientists and satirists everywhere must remain vigilant to protect liberties, and to fight obscurantism in any form. Social science and other research is needed to better understand the origins of violent fanaticism, conflict and intercommunal strife. Tackling terrorism is about much more than repressive measures. It demands long-term political and social initiatives, and policies to help to address the root causes.

The heritage of Voltaire and the Enlightenment explains why the French people have reacted much more strongly to the latest attacks than to the many acts of terrorism they have endured in the past. The terrorists attacked a symbol of the right to free expression. Free speech does have its limits, and many countries rightly impose laws that, for example, outlaw the incitement of religious or racial hatred. But the right to criticize, and even to mock, religion, fanaticism, superstition and indeed science is not only rightly protected by law in France, but is enshrined there, as in many countries, as a fundamental human right.

The killing also attacked other symbols of the Republic, including a kosher supermarket — one symbol of a multicultural society — and the police. One of the officers killed, Ahmed Merabet, was Muslim, and his brother aptly remarked that the killers were “false Muslims” and that Merabet had been “proud to represent the French police and to defend the values of the Republic — liberty, equality, fraternity”.

Marches for national unity on Sunday brought some 4 million people onto the country’s streets, the largest turnout in French history; the crowds in Paris surpassed those that welcomed allied troops’ liberation of the city during the Second World War. Fundamentalists throughout history have sought to subjugate freedoms, including freedom of expression and thought. More than 200 years ago, France rejected them with a “Non!” that echoed across the world. It has done so again. ■

# Deep mysteries

*Arguments among ocean scientists show how much remains to be discovered.*

It is the ultimate metaphor for the unknown: the deep; the depths; the abyss. The ocean is our ultimate ancestral home and a constant and alluring source of opportunity, danger and mystery.

Certainly, those scientists who spend their careers peering into the seven seas are quick to dredge up the idea that we know more about celestial bodies than we do about the oceans. Who can blame them when outer space grabs the attention? Witness the interest in the Rosetta comet mission or last week's announcement of more exoplanets — now counted in the hundreds — which received wall-to-wall press coverage.

News about the oceans tends to be bad, and feeds the dark imagery — such as Malaysia Airlines flight MH370, which is believed to have gone missing over the Indian Ocean with huge loss of life, or the enormous car-carrying ship that ran aground a fortnight ago off the UK coast.

Some scientists are now questioning whether this drip feed of negativity about the oceans is reflected in the attitudes of researchers. Carlos Duarte of the University of Western Australia in Crawley and his colleagues claim in a provocative paper that ocean science focuses too much on the narrative of man-made disaster (C. M. Duarte *et al. BioScience* <http://dx.doi.org/10.1093/biosci/biu198>; 2014). Their study argues that “doom and gloom media accounts” of the terrible state of the oceans are frequently not based on strong evidence.

Some marine situations painted as global calamities are well supported by science, they acknowledge, such as the depletion of fish stocks. But others are far from certain; for example, global explosions in jellyfish numbers and the predicted impact of ocean acidification on organisms such as corals. The media can be prone to exaggeration, but marine researchers, too, “may not have remained sufficiently skeptical”, say Duarte and his co-authors. Some ‘calamities’ are now accepted by researchers, and repeated as truth owing to problems with citations and observation bias, and to the desire to help solve these problems. This is

a contentious thesis, and other researchers dispute it. As usual in such spats, both sides argue that more work is needed. It is hard to disagree.

Some progress on one high-profile marine problem — the fate of coral reefs — is published on *Nature's* website this week. The study uses years of data from the Seychelles to unpick the factors that allowed coral reefs there to bounce back and flourish after a mass ‘bleaching’ event in 1998 (N. A. J. Graham *et al. Nature* <http://dx.doi.org/10.1038/nature14140>; 2015). Some 90% of the corals died and turned the reefs pale.

**“To neglect the oceans because it is cheaper to get good results on land is foolish.”**

Around half of the reefs recovered from the bleaching, but more worryingly, whether they were in a marine protected area seemed to make no difference to this recovery. Studies such as this will be increasingly important as climate change forces alterations in many marine systems. Even in well-studied ecosystems such as reefs, the effects of climate change and other human activity are unclear, and getting answers from such complex systems requires huge amounts of data. The Seychelles analysis used 17 years of data to come to its conclusions — data that were being collected when European scientists were still discussing parts of the Rosetta mission.

Some studies need even more data. Also published on *Nature's* website this week is a paper on sea-level rise that looks at tide data going back to the 1900s (C. C. Hay *et al. Nature* <http://dx.doi.org/10.1038/nature14093>; 2015). Such data have previously been used to estimate a sea-level rise of between 1.6 and 1.9 millimetres a year over the twentieth century. By contrast, Hay *et al.* find a lower rate of 1.2 mm per year. But their estimates for more-recent changes are 3 mm a year between 1993 and 2010 — in line with previous estimates. This implies that the recent increase in sea-level rise may be bigger than previously thought: a small but rather important difference.

Information on sea levels from some sites dates back to the eighteenth century. Yet the real value of this information is only now emerging. Ocean data are expensive to collect. Ships are costly to build, equally so to run. But to neglect the oceans because it is cheaper to get good results on land is foolish. The proverbial drunk searches for his keys under a streetlight because that is the only place that he can see. The sea may be mysterious, but some mysteries can be solved. ■

# Out of the bag

*The preference for either cats or dogs affects science more than you might think.*

The much-discussed difference between pet cats and dogs was neatly summarized by the late British journalist Christopher Hitchens. “Owners of dogs will have noticed that, if you provide them with food and water and shelter and affection, they will think you are god,” Hitchens observed. “Whereas owners of cats are compelled to realize that, if you provide them with food and water and shelter and affection, they draw the conclusion that they are god.”

Many see the apparent conflict between the attitudes of the two animals mirrored in the personalities of those who choose to own one or the other. Cat owners, it is claimed, are more neurotic and open with their emotions. Dog people are more disciplined and outgoing (and not just for long, muddy walks). Science has little conclusive to say on the matter, but accident statistics do: a 2010 study of non-fatal injuries in the United States found that more than seven times as many people were likely to hurt themselves in falls caused by dogs as by cats. (Cat owners are allowed a feeling of smug superiority here, because the actions of dog owners were as much to blame for the accidents in many cases as the animals themselves.)

As we discuss in a News story on page 252, the differences between dog folks and cat people extend all the way up to senior scientists, including some at the US National Institutes of Health, at least according to feline researchers. Work on cats has been overlooked for years, they complain, partly because “there were more powerful people interested in dogs”. The complete dog genome was sequenced a decade ago, and has produced hundreds of genes linked to canine traits and diseases, but a high-quality version of the first cat genome was published only last year.

Now, the cat lobby is trying to catch up, and feline fanciers everywhere have their chance to help. Just US\$7,500 will pay for a single cat's genome to be sequenced, and the funders — pet owners, breeders, pet-food companies — get to choose the breed or even the individual animal. Together, the project organizers hope that comparisons between dozens of these separate genomes will shed more light on cat diseases and genetic mutations that may drive similar conditions in humans — just as they have already for dogs.

Indeed, as scientists know, cats and dogs have more in common than it might seem. Cats and dogs do not even have to fight like, well, cat and dog. Plenty of people own both animals, and research on these households offers some advice: get the cat first, and get both while they are young.

➔ **NATURE.COM**  
To comment online,  
click on Editorials at:  
[go.nature.com/xhunqv](http://go.nature.com/xhunqv)

Don't fancy either? Then take inspiration from Winston Churchill. “I am fond of pigs,” he said. “Dogs look up to us. Cats look down on us. Pigs treat us as equals.” ■





## The focus on bibliometrics makes papers less useful

Forcing research to fit the mould of high-impact journals weakens it. Hiring decisions should be based on merit, not impact factor, says **Reinhard Werner**.

**H**ow do we recognize a good scientist? There is an entire industry — bibliometrics — that would have us believe that it is easy: count journal articles, sort them according to the impact factors of the journals, and count all the citations.

Science managers and politicians seem especially fond of such ways to assess 'scientific quality'. But many scientists also accept them, and use them in hiring and funding decisions. They are drawn to the alleged objectivity of bibliometrics. Indeed, one sometimes hears that scientists should be especially ready to apply scientific methods to their own output. However, scientists will also be aware that no good science can be built on bad data, and we are in a unique position to judge the quality of the raw data of bibliometrics, because we generate them through our citation behaviour.

The underlying assumption of bibliometrics is that, by citing, scientists are engaging in an ongoing poll to elect the best-quality academic papers. But we know the real reasons that we cite. Chiefly, it is to refer to results from other people, our own earlier work or a method; to give credit to partial results towards the same goal; to back up some terminology; to provide background reading for less familiar ideas; and sometimes to criticize.

There are less honourable reasons, too: to boost a friend's citation statistics; to satisfy a potential big-shot referee; and to give the impression that there is a community interested in the topic by stuffing the introduction with irrelevant citations to everybody, often recycled from earlier papers.

None of these citations — for good reasons or bad — express the opinion that the paper in question is a remarkable scientific achievement.

Consequently, highly cited papers often contain popular (but otherwise unimpressive) concepts or methods. If you have a favourite well-cited paper, it is a sobering experience to check 20 random citations. They typically contain little appreciation for the quality of the work.

To be sure, selection for an academic job guided mainly by citation statistics or papers in high-impact journals will get better results than flipping a coin. But it is blind to the difference between someone who creatively develops a research agenda — and is likely to be doing that in ten years — and someone who grinds out papers in a narrow, fashionable subfield.

Many negative effects of bibliometrics come not from using it, but from the anticipation that it will be used. When we believe that we will be judged by silly criteria, we will adapt and behave in silly ways.

A good example is the distortion in the journal landscape — and with it the changes in the style of papers — that arose when the journal impact factor began to be taken seriously as a proxy for reputation.

For example, when *Physical Review Letters* (PRL) split from *Physical Review*, it was intended to allow speedier publication of short announcements, which had previously been sent as unrefereed letters to the editor of *Physical Review*.

It is easier to reach high impact with this format, so the 'reputation' shifted from the standard journal to the letters section. Although there is no reason that shorter papers should be scientifically better than long ones, many authors now happily mutilate their work to stay under PRL's page limit, rendering papers less readable and less useful.

Another example is the way *Nature* became the top journal for experimental physicists. Life scientists are more numerous and use more citations than physicists, so the impact factors of *Science* and *Nature*, which cover all sciences, easily beat that of any non-review physics journal. Despite the higher impact factor, there is no reason why a paper written for a broad audience should be scientifically more valuable than one with an in-depth technical discussion. In fact, in pitching for such an audience, authors often leave out the tricky parts, keep technical terms out of their titles, and overstate their conclusions in broad terms.

What can we do? Simply, individual scientists must resist the trend of making bibliometrics a central plank in their decision-making processes. And we must make this public, perhaps by stating in job adverts that papers will be judged by scientific merit and not by journal impact factor.

Once a hiring decision is made, we should resist the temptation to justify it by quoting the candidate's bibliometrics to administrators. This reinforces the damaging idea that hiring decisions

could be made by administrators in the first place, and makes it harder to justify decisions that do not follow the metrics the next time round.

As the tyranny of bibliometrics tightens its grip, it is having a disastrous effect on the model of science presented to young researchers. For example, a master's student of mine moved to a renowned research institute for his PhD. Like many institutes, this one boasts of its performance in terms of publications in high-impact journals. So my student was told: "If you cannot write up your research in a form suitable for *Nature* or *Science* or *Physical Review Letters*, don't bother to even do it." Such advice, driven by the appeal of metrics to funders, is common but horribly misguided.

If we raise scientists to be driven by such extrinsic motivation alone, then why should they not follow the logic to its natural conclusion, and run off to become well-paid bankers instead? ■

**Reinhard Werner** is professor of theoretical physics at Leibniz University in Hanover, Germany.  
e-mail: reinhard.werner@itp.uni-hannover.de

WHEN WE BELIEVE  
THAT WE WILL BE  
JUDGED BY  
**SILLY  
CRITERIA**  
WE WILL ADAPT AND  
BEHAVE IN  
**SILLY WAYS.**

➔ **NATURE.COM**  
Discuss this article  
online at:  
[go.nature.com/q3edc3](http://go.nature.com/q3edc3)

# RESEARCH HIGHLIGHTS

Selections from the  
scientific literature

## MICROBIOLOGY

### Gut microbes' survival tactics

Gut bacteria protect themselves from host inflammation by modifying their outer membranes.

Immune responses designed to wipe out infection could, in theory, also perturb helpful flora that reside in the gut. To find out how these microbes resist the effects of inflammation, Andrew Goodman of Yale University in New Haven, Connecticut, and his colleagues studied 17 bacterial species that normally live in the human gut. They found that the microbes were all resistant to antimicrobial peptides released by hosts to kill pathogens.

In the bacterium *Bacteroides thetaiotaomicron*, this resilience was linked to expression of a protein called LpxF, which neutralized the negative charge of the cell membrane, preventing the positively charged peptides from binding to the gut microbe's surface. Mutants that did not express LpxF were outcompeted in mouse guts by other microbes during inflammation.

*Science* 347, 170–175 (2015)

## MATERIALS

### Arsenic forms a semiconductor

Single-atom-thick layers of arsenic and antimony could be efficient semiconductors that have more applications than other two-dimensional materials.

Atom-thick materials can have unique electronic and optical properties, but some operate only at certain wavelengths of light, owing to small 'band gaps'. On the basis of quantum mechanical calculations, Zhongfang



## ANIMAL BEHAVIOUR

### Monkey in the mirror

Macaques can be trained to recognize themselves in a mirror, the first such observation in any monkey species.

Most animals encountering their reflections act as if they are seeing another creature. To find out whether monkeys can be trained to recognize their own reflections, Neng Gong and his team at the Shanghai Institutes for Biological Sciences placed rhesus macaques in front of a mirror and shone a low-powered laser beam on their faces to produce a mild heat irritation. They rewarded animals when they touched the irritating spot on their face (pictured).

After 12 to 38 days of this regimen, 5 out of 7 macaques using the mirror touched an odourless mark applied to their faces. With a mirror in their cages, some of these monkeys seemed to use it to explore parts of their bodies that they could not otherwise see. It is not clear, however, whether these behaviours mean that the monkeys have higher cognitive abilities such as self-awareness, the authors say.

*Curr. Biol.* <http://doi.org/x54> (2015)

Chen at the University of Puerto Rico in San Juan, Haibo Zeng at the Nanjing University of Science and Technology in China and their colleagues predict that arsenic and antimony can switch

from being semi-metallic in bulk to semiconducting as a single-atom layer. These materials, called arsenene and antimonene, have wider band gaps than other two-dimensional semiconductors,

meaning that they could be used in short-wavelength optoelectronic devices such as blue or ultraviolet light-emitting diodes.

The authors say that such materials could soon be synthesized in the lab.

*Angew. Chem. Int. Edn*

<http://doi.org/f2x29z> (2015)

## EVOLUTION

### Lungs began with many chambers

The lungs of ancestral, land-based vertebrates may have had multiple chambers rather than just one, as was believed.

Markus Lambertz at the University of Bonn in Germany and his colleagues studied lung samples from 73 species of amniotes, which include mammals, birds and reptiles. They also looked at lung development in embryos of a gecko, *Paroedura picta*, which has single-chambered lungs. They found that all multi-chambered lungs shared key anatomical features, such as branching of the arteries. These features were present even in the single-chambered lungs of lizards and snakes, and in the embryonic gecko lung.

Ancestral amniotes evolved multi-chambered lungs as they shifted to life on land, the authors say. Some, however, may have later developed single-chambered lungs as they evolved into smaller creatures to maximize air space in the lungs, the team adds.

*Biol. Lett.* 11, 20140848 (2015)

## SUSTAINABILITY

### Resource use peaks worldwide

The rates at which humans consume multiple resources such as food and wood peaked at roughly the same time, around 2006. This means

NING GONG AND COLLEAGUES/CURR. BIOL.

J. RODGERS, UNIV. ILLINOIS

that resources could be simultaneously depleted, so achieving sustainability might be more challenging than was thought.

Ralf Seppelt of the Helmholtz Centre for Environmental Research in Leipzig, Germany, and his colleagues estimated the peak rate of extraction for 27 resources. For 20 of them, mostly renewables such as meat and rice, the peak-rate years occurred between 1960 and 2010, with many clustering around 2006. Only coal, gas, oil, phosphate, farmed fish and renewable energy have yet to peak.

Humans use multiple resources to generate new ones and to meet basic needs, which could explain the synchronicity of peak usage, the authors suggest.

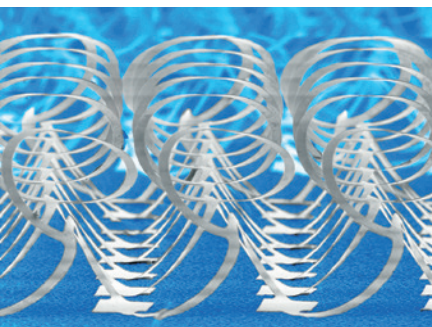
*Ecol. Soc.* 19, 50 (2014)

#### MATERIALS

## Silicon buckles to form 3D shapes

Researchers have created a variety of small, three-dimensional structures by buckling strips of silicon and other materials.

Turning advanced two-dimensional materials into three-dimensional shapes has proved difficult. John Rogers at the University of Illinois at Urbana-Champaign and his colleagues added hydroxyl groups at specific locations along the length of silicon ribbons as narrow as 800 nanometres and as thin as 100 nm. They also added these groups in specific patterns to an elastic, stretched substrate, and allowed the silicon strips to bond to the substrate.



When the substrate returned to its original shape, the silicon buckled to form a range of structures less than a millimetre wide (**pictured**), including helices, boxes and flowers. These could then be assembled into larger configurations.

The method uses a variety of materials, such as metals and polymers, and has the potential to make structures for a wide range of electronic devices, the authors say.

*Science* 347, 154–159 (2015)

#### EVOLUTION

## Mosquitoes gain resistance

A malaria-carrying mosquito inherited insecticide-resistance genes from a related species, around the time that bed nets treated with insecticide were increasingly used in West Africa.

Gregory Lanzaro at the University of California, Davis, and his colleagues analysed DNA from more than 1,000 specimens of *Anopheles coluzzii* and *Anopheles gambiae* in Mali from 2002 to 2012. They found that a group of genes, including one for insecticide resistance, from *A. gambiae* moved into *A. coluzzii* around 2006 when the two species mated.

Campaigns to encourage the use of insecticide-treated bed nets began in 2005 in this region, and the authors suggest that the nets favoured the selection of hybrid, insecticide-resistant mosquitoes.

*Proc. Natl Acad. Sci. USA* <http://doi.org/x56> (2015)

#### PHOTONICS

## Few photons make 'ghost image'

Physicists have captured an image of a wasp's wing using less than one photon per pixel.

Peter Morris and his colleagues at the University of Glasgow, UK, used a technique called ghost imaging, which uses pairs of photons whose positions

## SOCIAL SELECTION

Popular articles on social media

### Unveiling secret funding decisions

Many scientists struggle to understand why some grant applications succeed and others fail, perhaps explaining the online popularity of an article calling for increased transparency in the grant peer-review process. Writing in *PLoS Biology*, Daniel Mietchen, an evolutionary biologist at Berlin's Museum of Natural History, argues that, among other things, all successful proposals and their reviews should be released to the public. "Open science is now old hat; here's a call for open research funding processes," tweeted Dorothy Bishop, a neuropsychologist at the University of Oxford, UK. But some see a downside to transparency. James Johnson, a diabetes researcher at the University of British Columbia in Vancouver, Canada, tweeted: "I try to be open, but I've also had desperate people steal my ideas." *PLoS Biol.* 12, e1002027 (2014)



Based on data from altmetric.com. Altmetric is supported by Macmillan Science and Education, which owns Nature Publishing Group.

**NATURE.COM**  
For more on popular papers: [go.nature.com/tm0blc](http://go.nature.com/tm0blc)

are inextricably correlated, or entangled. One of each photon pair was either transmitted or absorbed by the wing, while its twin was used to reconstruct the image.

Each photon was collected at a separate detector. To avoid recording stray photons that cause noise, a camera captured the 'image-creating' photon only when its partner from the wing was detected. Applying image-compression techniques, the authors further reduced the number needed to just 0.45 photons per pixel.

Such techniques could be useful in biological imaging when high levels of light could damage the sample, they say. *Nature Commun.* 6, 5913 (2015)

#### PLANT GENETICS

## Maize's journey out of Mexico

DNA from the cobs of ancient maize (corn) shows how the crop was taken to the US southwest from Mexico.

Maize was domesticated from the wild grass teosinte, an inedible weed, more than

6,000 years ago in southern Mexico, and later spread throughout the Americas. A team led by Rute da Fonseca and Thomas Gilbert at the University of Copenhagen analysed nuclear DNA from 32 maize samples from several archaeological sites spanning Mexico and the US southwest (**pictured** is a 5,000-year-old specimen).

They conclude that maize arrived in the US southwest around 4,000 years ago along a highland route in central Mexico — not by a Pacific coastal route as other studies had suggested. Along the way, maize evolved into a sweeter and more drought-tolerant crop.

*Nature Plants* <http://doi.org/x6p> (2015)

**NATURE.COM**  
For the latest research published by Nature visit: [www.nature.com/latestresearch](http://www.nature.com/latestresearch)





# SEVEN DAYS

The news in brief

## FACILITIES

### Netting neutrinos

Officials in China broke ground at the Jiangmen Underground Neutrino Observatory (JUNO) in Jinji Town on 10 January. The underground experiment, an international collaboration designed to succeed China's Daya Bay Reactor Neutrino Experiment, is scheduled to begin taking data in 2020. JUNO aims to shed light on the relative masses of the three known types of neutrino by studying particles emitted by nearby nuclear reactors or coming from space. Earlier last week, India's government approved the India-based Neutrino Observatory (INO), at an estimated cost of US\$240 million.

## POLICY

### Pipeline gets closer

On 9 January, the Nebraska Supreme Court dismissed a lawsuit that had sought to block the planned route of a controversial oil pipeline. The Keystone XL pipeline is intended to transport up to 830,000 barrels of crude oil per day from the tar sands of western Canada to the US Midwest, where it would connect with existing pipelines to refineries along the Gulf of Mexico. US President Barack Obama has not said yet whether he will approve the project, but on 7 January his administration said that he plans to veto legislation introduced by Republicans in Congress that would circumvent an ongoing review and fast-track construction.

### Nuclear disposal

The private consortium charged with cleaning up a huge nuclear-waste repository in Sellafield, UK, has lost its contract, the UK

government announced on 13 January. Last year, a cross-party group of politicians complained that costs had increased at the site and that targets for the reprocessing of waste had been missed. Costs on one section of the repository alone increased by more than £300 million (US\$450 million) between March 2012 and September 2013, while estimated completion dates on another project were put back by years. Control of the company set up to manage the clean-up will now switch from the private consortium to the UK Nuclear Decommissioning Authority.

## EVENTS

### Philae still missing

The search for Philae, the robotic lander that had a bouncy touchdown on a comet on 12 November, has become more difficult. The European Space Agency lost contact with the probe after it came to rest in a shaded spot away from its planned landing site and went into hibernation after its solar-powered batteries ran out. Rosetta, the spacecraft that deployed Philae, is currently in a higher orbit around comet 67P/Churyumov–Gerasimenko than it was in December, meaning that its photographs of the comet's

surface are fuzzier. Mission scientists hope to perform a fly-by, which would afford Rosetta a better view.

### SpaceX rocket crash

The first attempt to land a reusable rocket on a boat has ended in a crash. The private company SpaceX, of Hawthorne, California, successfully launched an unmanned resupply ship to the International Space Station on 10 January — but failed to land the first stage of its Falcon 9 rocket on a floating platform in the Atlantic Ocean as planned because the fins that helped to stabilize its descent ran out of hydraulic



ALFRED ROSENBERGER

## Divers find ossuary in Madagascan cave

A giant trove of bones of extinct lemurs and other animals has been found in an underwater cave in Madagascar's Tsimanampesotse National Park, researchers announced last week. The specimens recovered so far include the 30-centimetre-long head of a massive 'koala lemur' (*Megaladapis*) and dozens of giant lemurs (*Pachylemur insignis*), but thousands of bones may still lie on the silty floor. Lemurs are unique to the island country, and the

anthropologists estimate that two-thirds of the species that existed only a thousand years ago are now extinct. The cave, which can only be reached by skilled divers, also contains remains of other large extinct animals, including elephant birds (*Mullerornis*), giant fossa (*Cryptoprocta spelea*), horned crocodiles (*Voay robustus*, pictured) and pygmy hippos (*Hippopotamus lemerlei*). The team hopes to return to retrieve more specimens this summer.

fluid. SpaceX is one of two companies tasked with flying cargo to the space station for NASA. Among other supplies, the ship carried an instrument to measure particulates in the atmosphere (see *Nature* <http://doi.org/x83>; 2014).

## Parasite in decline

A global campaign to eradicate guinea-worm disease is getting closer to its goal. The Carter Center, a medical charity in Atlanta, Georgia, announced on 12 January that only 126 cases of the disease were reported in 2014, a 15% drop from the previous year. All of the 2014 cases occurred in one of four African countries. The majority — 70 cases — were in South Sudan. The Carter Center said that when it first launched its campaign in 1986, 3.5 million people were afflicted with the tropical disease every year in Africa and Asia.

### PEOPLE

## Policy leader dies

Hubert Markl (pictured), one of the most influential figures in German research policy, died on 8 January, aged 76. Markl was president of the DFG, West Germany's major funding agency for basic research, during the reunification with East Germany in 1990. From 1996 to 2002, he led the Max Planck Society, a leading



research organization, through a modernization effort. He closed poorly performing institutes in the country's west and opened new ones in the east, often with prominent foreign scientists as directors. Markl also launched an independent study of the society's activities during the Third Reich and formally apologized for the society's Nazi-era crimes.

### FUNDING

## Ebola vaccine trials

Clinical trials of Ebola vaccines could begin in West Africa within a few weeks, officials of the World Health Organization in Geneva said after a high-level meeting on 8 January. The trials will test the efficacy of up to three different vaccines at preventing infection in people in Guinea, Liberia and Sierra Leone. On 12 January, the

Wellcome Trust, a London-based biomedical-research charity that is involved in the trials, released a 'road map' on Ebola vaccines, emphasizing the need to involve local communities in organizing the trials.

## Stimulating Japan

The Japanese government approved a ¥3.1-trillion (US\$26-billion) supplementary budget on 9 January, to jolt the economy. The government will spend ¥351.5 billion to boost energy efficiency in the industrial sector and ¥5.2 billion on research related to developing energy sources. The package reportedly included ¥1.2 billion for efforts to develop next-generation power semiconductors at Nagoya University. Another ¥3.7 billion will support increased monitoring of Japan's volcanoes.

### BUSINESS

## Genome deals

The personal-genetics start-up 23andMe of Mountain View, California, has entered the first two of what it promises will be a series of agreements with major pharmaceutical companies. On 6 January, Genentech of South San Francisco, California, disclosed that it will pay up to US\$60 million to access part of

23andMe's database containing the genome sequences of 3,000 people with Parkinson's disease. And on 12 January, New York City-based Pfizer announced a broad agreement to access the collected genetic information of 800,000 23andMe customers. Both deals were seen as motivated by the companies' interest not only in 23andMe's genetic data, but also in the personal data that its customers share with 23andMe, which could accelerate and improve clinical trials.

## Pharma deal

Swiss pharmaceutical giant Roche announced on 12 January that it will spend more than US\$1 billion to purchase more than half the shares of biotechnology firm Foundation Medicine. The firm, in Cambridge, Massachusetts, generates genetic profiles of cancer tumours with the aim of finding treatments that can be personalized to patients. Roche, headquartered in Basel, says that it wants to sell Foundation's DNA tests outside the United States, and will work with the firm to create tests that use blood samples, rather than a tissue biopsy.

## Diversity push

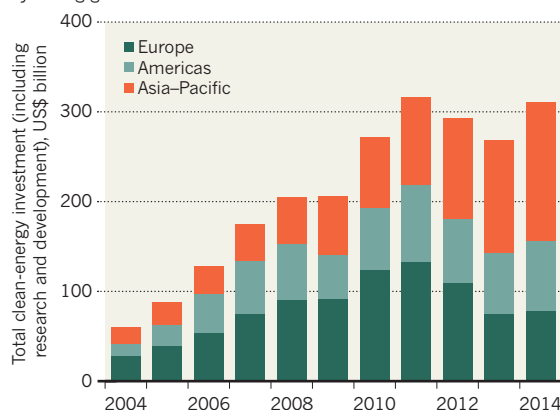
Computer-chip giant Intel on 6 January pledged to spend US\$300 million by 2020 to boost the diversity of the technology industry's workforce. The company plans to invest in recruiting and retaining engineers and computer scientists who are women or from under-represented ethnic groups and to expand its educational programmes at schools and universities. Chief executive Brian Krzanich said that Intel aims to raise the representation of women and minorities at the company, in part to better cater to Intel's diverse customer base.

## TREND WATCH

Global investment in clean-energy projects and research surged in 2014, according to figures released on 9 January by Bloomberg New Energy Finance. Almost half of spending was on solar power, which rose by 25% from 2013 to US\$149.6 billion. Clean-energy investment in China, the sector's leading spender, grew by 32% to a record \$89.5 billion, whereas spending in Europe rose by only 1% to \$66 billion. The recent slump in the prices of oil and natural gas is expected to slow global investment in renewables.

## CLEAN-ENERGY INVESTMENT REBOUNDS

Global spending on renewables grew by 16% in 2014, boosted by strong growth in Asia.





# NEWS IN FOCUS



**PET SCIENCE** Cats catch up to dogs in the genome-sequencing race **p.252**

**ENVIRONMENT** Digital island built for eco-experimentation **p.254**

**SPACE** 'GoreSat' resurrected after 14 years in limbo **p.256**

**SQUID SCIENCE** The bacteria that serve as live-in light bulbs **p.262**

UESLEI MARCELINO/REUTERS/CORBIS



Brazilian president Dilma Rousseff (left) has chosen climate-change sceptic Aldo Rebelo (right) to lead her science ministry.

## ENVIRONMENT

# Political appointments spur concerns for Amazon

*Environmentalists worried after Brazilian president picks ministers with ties to agriculture lobby.*

BY JEFF TOLLEFSON

Brazilian president Dilma Rousseff survived an unexpectedly strong challenge from a prominent environmentalist to stay in office. Now, less than three months into her second term, Rousseff has sparked controversy by appointing an avowed climate-change sceptic as science minister.

The choice of Aldo Rebelo, a Communist law-maker who last July called the idea of global warming “incompatible with current

knowledge”, follows a decade in which Brazil has slowed Amazon deforestation and claimed a leadership role in international climate policy. Rebelo promised to uphold the government’s climate stance when he assumed his post at the Ministry of Science, Technology and Innovation on 2 January, but his appointment has stirred fears that Brazil may backslide on environmental protection.

“In general, scientists are sceptical and disappointed,” says Jean Ometto, coordinator of the Earth System Science Centre at the

government’s National Institute for Space Research in Sao Jose dos Campos. “But the scientific centre is strong, and there’s going to be pressure from the scientific community not to change the way that climate change is governed.”

Rebelo earned his reputation as a skilful politician in the lower house of Brazil’s National Congress, and he played a leading part when the ‘*ruralistas*’ — a voting bloc representing rural agriculture — weakened protections for Brazil’s forests in 2012. Most scientists say that Rebelo largely ignored the scientific ►



► community during that debate, although some think that he may move to the political centre as minister of science.

In a written statement provided to *Nature*, Rebelo dismissed concerns about his position on global warming. “The debate about climate change exists independent of my opinion,” he said. More broadly, Rebelo said that his duty as minister of science is to listen to — and fight for — the academic and scientific community in Brazil.

He has limited power to influence climate policy, although the ministry does oversee some national assessments and some climate reports to the United Nations. One area that will receive considerable attention is the ministry’s funding for science, which has been squeezed by competing demands in recent years.

But Rebelo is not the only appointment to raise eyebrows. Rousseff named Kátia Abreu, who fought environmentalists as a senator and agricultural lobbyist, to head Brazil’s agriculture ministry. And both appointments come at a sensitive time: although the rate of Amazon deforestation has dropped by 82% since 2004, the pace of land clearing spiked sharply in 2013 — and it seems to be on the rise once again.

Rousseff has never had a good relationship with environmentalists, who say that she is hiding behind Brazil’s relative success in reducing deforestation while pushing for major dams and ports that will only increase pressure on forests. Now, after a bruising election, she is looking to rebuild support in Congress in the middle of corruption scandals that have tainted her party.

The Communist party, of which Rebelo is a member, has been a “small but loyal” member of the governing coalition led by Rousseff’s Workers’ Party, and it is also friendly with the agricultural caucus, says Steve Schwartzman, an anthropologist with the Environmental Defense Fund in Washington DC, who has worked in Brazil for decades.

One key to Brazil’s success in curbing deforestation and carbon emissions over the past decade was the involvement of multiple agencies, but that kind of coordinated approach could be more difficult today, Schwartzman says. “These ministerial appointments cast a lot of doubt on what we can expect from this government.”

Helena Nader, president of the Brazilian Society for the Advancement of Science, takes a more conciliatory approach: she says that scientists should give Rebelo some time at the ministry before judging him. More broadly, Nader says that she isn’t expecting a major backslide on environmental issues by the new Rousseff government.

“Nevertheless,” she says, “only time will show.” ■



The cat genome is out of the bag, and has already helped to pinpoint a gene involved in kidney disease.

#### GENOMICS

## ‘I can haz genomes, too’

*Cats sparked an Internet meme complete with its own grammar, but dogs have dominated genetics labs — until now.*

BY EWEN CALLAWAY

Cats may have beaten dogs on the Internet but felines have been a rare breed in genetics labs compared with their canine counterparts. Now, at last, cats are clawing their way into genomics.

At a meeting this week in San Diego, California, a close-knit group of geneticists unveiled the first results from an effort to sequence the genomes of 99 domestic cats. The work will benefit both humans and felines, the researchers say, by mapping the mutations underlying conditions that afflict the two species, such as kidney disease.

“It’s a great time to be in cat genomics,” says William Murphy, a geneticist at Texas A&M University in College Station who is involved in the effort. Plummeting costs for DNA sequencing now make it possible to do genomics cheaply — and cat genomics, long under-funded compared with similar efforts in dogs, is benefiting, he says. “We’re finally at the point where we can do all sorts of things we wanted to do 5 or 10 years ago.”

The first cat genome sequence<sup>1</sup> — from an

Abyssinian named Cinnamon — was reported in 2007. But the sequence contained significant gaps and errors, which slowed efforts to map genes. A high-quality version of Cinnamon’s genome was not published<sup>2</sup> until late 2014. Domestic dogs, meanwhile, have become a darling of geneticists: their full genome<sup>3</sup> was reported in 2005, and the sequence has been continually improved. Hundreds of genes underlying canine diseases and traits are estimated to have been discovered, compared with as few as a dozen for cats.

The discrepancy can be traced back to the early 2000s. After the completion of the human, mouse and rat genomes, the US National Institutes of Health organized a commission to decide on their next target; the dog genome was selected for high-quality sequencing, whereas cats were put on hold.

That got some cat geneticists’ backs up. “The truth is there were more powerful people interested in dogs,” says Stephen O’Brien, director of the Theodosius Dobzhansky Center for Genome Bioinformatics in St Petersburg, Russia, who led the initial cat-sequencing efforts.

But canine researchers were able to make a

compelling case. Pet dogs suffer from many of the same conditions as humans, from narcolepsy to arthritis. And the intensely inbred nature of dog breeds made it relatively easy to identify disease-causing genes: because there is little genetic variation within any particular breed, the genes that cause disease in affected individuals stand out.

Dogs had other advantages, too. The existence of kennel clubs, which maintain 'breed standards' and are full of enthusiastic pet owners and veterinary surgeons, helped dog geneticists to recruit subjects for study. "Given the resources they had, they were discovering new genetic diseases in breeds almost daily," says Niels Pedersen, a veterinary scientist at the University of California, Davis.

In fact, both cats and dogs offer insights into human disease, including those associated with old age. In 2004, a team led by geneticist Leslie Lyons of the University of Missouri in Columbia (and owner of two female cats, Withers and Figaro) discovered that mutations that cause polycystic kidney disease — a major cause of renal failure in older individuals — occur in the same gene in humans and cats<sup>1</sup>. Cat versions of type 2 diabetes, asthma, retinal atrophy and numerous other conditions have close similarities to human disease. Cats can also become infected with a virus that is closely related to HIV and experience symptoms similar to those of people with AIDS.

In the hopes of speeding up the discovery

of genes related to these conditions last year, Lyons launched the 99 Lives cat genome sequencing initiative, playfully hosted on a site called Lyons' Den. She discussed the effort on 11 January at the Plant & Animal Genome conference in San Diego.

Lyons' team is cobbling together funding from anywhere it can find it. The researchers are asking private owners, breeders and

**"I would love to eradicate all genetic disease in cat breeds before we're done."**

Any owner can participate," she says. All the data will be made public after the results are published.

With the money raised so far, the team has sequenced the genomes of 56 cats, including fancy breeds such as Burmese; cats with specific diseases; and a kitten named Dragon and his parents Ares and Marcus — the hope is to use the feline trio to narrow down the genetic basis for traits they share, such as their silver, curly coats.

Even Robert Wayne, a canine geneticist at the University of California, Los Angeles, agrees that Lyons' effort is important. "I hope she raises money for it," he says.

Insights from cat genomics extend beyond

even pet-food companies to donate the US\$7,500 needed to sequence the genome of a single cat, which could be one of a donor's choice. "Any cat can participate.

disease. Razib Khan, an evolutionary geneticist at the University of California, Davis, wants to use genome sequences to chart the domestication and spread of cats throughout the world, and to determine how different domestic and wild cats are genetically. "There's always the question — are they domesticated at all?" he says. The 2014 publication that included Cinnamon's genome already identified differences between domestic and wild cats, including genes expressed in the brain that are possibly linked to the docility of (some) house cats. "Wild cats will hand you your behind if you get next to them and domestic cats will sit on your lap," O'Brien notes.

Lyons is also keen to see genomics help felines. "I would love to eradicate all genetic disease in cat breeds before we're done," she says. Her team's discovery of the cause of polycystic kidney disease has reduced its prevalence among Persians, by removing cats with the mutation from the breeding pool. Her lab is now developing drugs that could treat the terminal condition in cats — and perhaps in humans. But human health, Pedersen says, is not the only goal. "I'm in it, and Leslie's in it, for the good of cats." ■

1. Pontius, J. U. *et al. Genome Res.* **17**, 1675–1689 (2007).
2. Montague, M. J. *et al. Proc. Natl Acad. Sci. USA* **111**, 17230–17235 (2014).
3. Lindblad-Toh, K. *et al. Nature* **438**, 803–819 (2005).
4. Lyons, L. A. *et al. J. Am. Soc. Nephrol.* **15**, 2548–2555 (2004).

## HEALTH

# First biosimilar drug set to enter US market

*But such cheaper, generic versions of biological drugs face scientific, regulatory and patent hurdles.*

BY HEIDI LEDFORD

**A**fter years of debate, the US Food and Drug Administration (FDA) is poised to allow the sale of biosimilars, cheaper versions of complex and expensive biological drugs used to treat conditions such as cancer and autoimmune diseases. On 7 January, an FDA advisory panel decided unanimously that a drug made by Sandoz, the generics arm of Swiss pharmaceutical giant Novartis, should be accepted as a replacement for filgrastim (Neupogen), an immune-boosting drug for people undergoing cancer treatment made by Amgen of Thousand Oaks, California.

Such knock-offs are called biosimilars

because the drugs that they mimic, dubbed biologics, consist of complicated molecules that are made in living cells and are impossible to copy exactly. Even copying them inexact is immensely challenging — despite the expected approval of the Sandoz drug, the difficulties involved in creating and evaluating biosimilars may limit their infiltration of the marketplace. The field is also littered with patent issues, especially with regard to how the drug is manufactured. In the case of filgrastim, Sandoz is challenging some of the legal requirements for approval.

"We're starting from scratch," says Jordan Paradise, a specialist in technology law at Seton Hall University in Newark, New Jersey. "A lot of

the scientific uncertainty is still there."

Unlike typical drugs, which are relatively small molecules made through biochemical processes, biologics are large protein molecules produced by genetically engineered organisms. Living cells may chemically modify the proteins they make by adding complex sugars and other compounds at certain positions. The exact conditions under which cells are grown can alter the pattern of these modifications, and thus the molecule's structure and behaviour. The result is a drug so complex that it is difficult — if not impossible — to fully characterize.

Because biosimilars are inexact copies, they are required to undergo more testing than an ordinary generic drug. The European Union has been evaluating and approving biosimilars for the past decade, but the United States did not have a way to do so until regulatory legislation was passed in 2010. Biotechnology companies have been waiting to find out how the FDA would evaluate the drugs.

Patient advocates hope that biosimilars can reduce drug costs by increasing competition. Biologics are expensive: researchers have calculated that treatment of metastatic colorectal cancer with bevacizumab (Avastin) costs about US\$75,000 per year of life gained (V. Shankaran *et al. Oncologist* **19**, 892–899; 2014). A report last year by the RAND ▶



► Corporation in Santa Monica, California, estimated that biosimilars could save \$44.2 billion by 2024.

Filgrastim is relatively simple: it is a small protein with no attached sugars. Even so, Sandoz presented the FDA with clinical-trial data from 388 people with breast cancer and 174 healthy participants to show that its biosimilar breaks down in the body similarly to the original, and does not provoke an immune response.

The FDA is expected to make a final decision by May. But even as Sandoz prepares to sell its drug in the United States, it is embroiled in a patent fight with Amgen. By US law, Sandoz had to reveal the details of its manufacturing method to Amgen — a provision not present in Europe — so that Amgen could determine whether any of its patents had been violated. Sandoz refused. That is a disheartening precedent, says Paradise. “Here we’ve got the first biosimilar application and we’ve already got the

manufacturers not working together.”

Such concerns loom large among manufacturers, says Nicholson Price, a patent-law specialist at the University of New Hamp-

**“We’ve got the first biosimilar application and we’ve already got the manufacturers not working together.”**

shire School of Law in Concord. Drug firms often keep their manufacturing methods confidential, and production of complex drugs gives them ample opportunity to file patents on manufacturing techniques or ways of characterizing molecules. “The second company is attempting to feel its way in the dark to what the first company has done,” says Price. “I suspect there are other biosimilars that are being deterred either by specific patents or just the worry that there may be

patents lurking out there that they don’t know about.”

Even when a biosimilar makes it over these hurdles, it is unclear how consumers will react to a drug that is almost, but not quite, a copy of the original. At the advisory-committee meeting, a number of patient groups expressed support for biosimilars and the promise of relief from high drug prices. But they voiced concerns that biosimilars would be given the same generic names as the drugs they were meant to replace, creating confusion as to whether recipients were getting the original or the copy. Many will be watching the Sandoz drug’s approval to see whether the FDA will let it be called filgrastim.

Committee member James Liebmann, an oncologist at the University of Massachusetts in Worcester, reacted to that concern with surprise. “This has been pretty clearly shown to be filgrastim,” he said. “To name it anything else would be misleading.” ■

## MICROSCOPY

# Inflated brains show nano detail

*Diaper material expands tissue, enabling ordinary microscopes to reveal nanoscale features.*

BY EWEN CALLAWAY

**A**n innovative method could enable biologists to image an entire brain in exquisite molecular detail using an ordinary microscope.

The technique, called expansion microscopy, involves physically inflating biological tissues using a material more commonly found in baby nappies, or diapers. Edward Boyden, a neuroengineer at the Massachusetts Institute of Technology (MIT) in Cambridge, discussed the technique, which he developed with his MIT colleagues Fei Chen and Paul Tillberg, at a conference last month.

Conventional optical microscopes cannot distinguish objects that are closer together than about 200 nanometres. Although super-resolution microscopy can discern objects as close together as about 20 nm, they require expensive, specialized equipment, and struggle with thick structures such as sections of brains.

“What we’ve been trying to do is figure out if we can make everything bigger,” Boyden told the meeting at the US National Institutes of Health (NIH) in Bethesda, Maryland. To do this, his team used a chemical called sodium acrylate, which has two useful properties: it can form a dense mesh that holds proteins in place, and it swells in the presence of water. The acrylate is the same substance that gives nappies their sponginess. When inflated, Boyden’s tissues grow by a factor of about 4.5 in each dimension.

Before swelling, the tissue is treated with a chemical cocktail that makes it transparent, and then with fluorescent molecules that anchor specific proteins to the acrylate, which is then infused into the tissue. Just as with nappies, adding water causes the acrylate to swell. After stretching, the fluorescent-tagged molecules move farther away from each other; proteins that were previously too close to distinguish with a visible-light microscope come into crisp

focus. In his presentation, Boyden suggested that the technique can resolve molecules that are as close as 60 nm before expansion.

Crucially, the process generally maintains the relative orientation and interconnection of proteins and keeps other cellular structures intact: it distorts the relative position of proteins by just 1–4%, Boyden’s team calculated.

Viviana Gradinaru, a neuroscientist at the California Institute of Technology in Pasadena, says that Boyden’s technique is another example of how scientists are bypassing hardware limitations by modifying biological tissue.

“This is certainly highly ingenious, but how much practical use it will be is less clear,” notes Guy Cox, a microscopy specialist at the University of Sydney, Australia. “If this is to be of any serious use, I suspect it will be in collaboration with existing super-resolution techniques on small macromolecular complexes, to push the boundaries a bit further, rather than looking at whole cells.” ■



**MORE  
ONLINE**

## TOP NEWS



Game theorists crack poker  
[go.nature.com/tgmtrq](http://go.nature.com/tgmtrq)

## MORE NEWS



● Macaques show self-recognition  
[go.nature.com/ipo7xu](http://go.nature.com/ipo7xu)

## NATURE PODCAST



Brain cells hibernate, climate displaces people, and *Nature Plants* debuts [nature.com/nature/podcast](http://nature.com/nature/podcast)



## DATA HEAVEN

Moorea is one of the most studied ecosystems in the world. Myriad data collected over four decades will be used to build a digital replica of the island that includes its varied geography, its climate and all of its plant and animal life.

**Social science**

Government census data will be combined with tourist numbers, employment status and economic revenue.

**Peaks and valleys**

Landscape model to 70-centimetre resolution created from satellite images.

**Ocean circulation**

Time-series measurements of currents, waves and water properties from an underwater sensor array around the island.

**Catalogue of life**

DNA barcodes for every species more than 1 millimetre in length.

**Watery secrets**

Continual sampling of the water's temperature, salinity, pH and microbial diversity at sites around the island, including the Tiahura Marine Protected Area, shown here.

**Coral reefs**

Long-term trends in coral and fish populations, including numbers and species composition.

**Underwater terrain**

Sea floor mapped using satellite imagery and sonar data from ships. Will have a resolution ranging from 0.5 metres in the shallows to 20 metres in the deep ocean.

## ENVIRONMENT

# Tropical paradise inspires virtual ecology lab

*Digital version of Moorea will provide a way to experiment with an entire ecosystem.*

BY DANIEL CRESSEY

A paradise on Earth could soon become the first ecosystem in the world to be replicated in digital form in painstaking detail, from the genes of its plants and animals to the geography of its landscape.

An international team is preparing to create a digital avatar of the Pacific island of Moorea, which lies off the coast of Tahiti and is part of French Polynesia. Moorea is already one of the most studied islands in the world; the team plans to turn those data into a virtual lab that would allow scientists to test and generate hypotheses about the impact of human activities.

Ecologists have used models for years to tease out the relationships between different facets of nature, such as temperature and population or predators and prey. But much of that modelling is relevant only to specific species or

research questions, and some scientists want a holistic view. As human activity and natural variations combine to alter the environment, researchers need to know how mitigating steps — such as setting up protected areas, or attempts to curb fossil-fuel use — might affect an entire ecosystem.

“We know the world’s changing. Yet the decisions we’re making, we’re making them in the dark,” says Neil Davies, one of the people behind the Moorea IDEA (Island Digital Ecosystem Avatars) project and director of Gump Station, the University of California, Berkeley’s marine-science base on the island. “We’re not going to have precise predictions ever, but we need to have a way of modelling different scenarios.” For example, if a hotel is built at a certain location, how does that change the ecosystem? If a species disappears from a river, what happens downstream?

Moorea is an ideal place to start, says Davies, because the island is about 16 kilometres across and has just 17,000 people living on it, making it easier to model than larger ecosystems and those that are more connected to the rest of the world. In addition, French researchers have been there since the 1970s, and Gump Station has been operating since the 1980s. Both efforts have collected myriad data on the island’s waters, with decades-long studies of coral and fish numbers (see ‘Data heaven’).

These traditional surveys of marine life are now being linked up with the Moorea Biocode Project, which aims to characterize every species larger than a millimetre in length on the island and allocate them a ‘DNA barcode’ — snippets of DNA that can be used as a unique identifier. Species can thus be identified quickly and easily even when they are in places or states that would otherwise be difficult ▶

► to recognize, such as in the contents of another organism's stomach, or in seed or larval form.

The avatar would combine insights gleaned from the Biocode project — such as which species are present at certain ocean spots, or which species are eaten by another — with data on weather, ocean currents and society such as population density and real-estate prices. It would provide a three-dimensional visualization of the island and its surrounding waters that might look something like those on Google Earth, but would enable researchers to zoom into a location, access data and run simulations.

"The first stage will be a framework to integrate the data we have. To collate them, combine them, and to make the data accessible to scientists," says project member Matthias Troyer, a computer scientist at the Swiss Federal Institute of Technology in Zurich. "Then, based on that, one can start on modelling."

#### EXPANDING PROJECT

The IDEA project was born in 2013, the brainchild of Davies, Troyer and three other marine scientists: Dawn Field at the University of Oxford, UK; Sally Holbrook at the University of California, Santa Barbara; and Serge Planes from the French research base on Moorea. The consortium now has more than 80 participants.

At meetings late last year, the IDEA team discussed how to combine existing data with those coming from the latest technologies. Some of the framework for the avatar is already under construction, and Davies says that the team is seeking funding of around US\$5 million over three years to pursue a pilot project.

The project is "really novel in the modelling community", says Mike Harfoot, an ecosystem modeller at the United Nations Environment Programme's World Conservation Monitoring Centre in Cambridge, UK, because it will integrate societal data with physical and biological components. And, he adds, the computational power required to take a holistic approach to modelling ecosystems has only recently become available.

"It's impressive the amount of data going into it," says Rick Stafford, a computational ecologist at Bournemouth University, UK. Getting the different data sets to talk to each other will be a challenge, but the time is ripe for such an ambitious undertaking, says Davies. And it if works on Moorea, the approach could be rolled out to other parts of the world. Although ambitious, says Davies, "it's not a pipe dream". ■

**"It's impressive the amount of data going into it."**



The DSCOVR craft is removed from a container in Cape Canaveral, Florida.

#### SPACE

# Mothballed NASA craft to launch

*Proposed by former US vice-president Al Gore in 1998 to image Earth, DSCOVR probe will monitor space weather.*

BY MARK ZASTROW

**A**fter nearly 14 years in limbo, an Earth-monitoring spacecraft built by NASA is finally set to fly.

The Deep Space Climate Observatory (DSCOVR), scheduled to launch as soon as 29 January, will constantly observe Earth's sunlit side from a distance of 1.5 million kilometres. It will track daily weather patterns and seasonal vegetation changes, monitor atmospheric pollution and make the most precise measurements yet of how much energy Earth throws out into space — crucial data for the improvement of global climate models.

DSCOVR's resurrection is thanks to renewed interest in what was originally its secondary mission: monitoring space weather. From a point between the Sun and Earth at which the bodies' gravitational pulls cancel out, the probe will be able to detect approaching solar storms — bursts of charged particles and powerful radiation that pose a threat to astronauts, orbiting satellites and power grids on the ground. Such storms are of interest to the US Air Force, which is funding the satellite's launch, and the National Oceanic and Atmospheric Administration (NOAA), which will operate it.

DSCOVR was the brainchild of former US vice-president Al Gore. He imagined a probe that would beam down a live image of Earth's illuminated side that could be available online.

Just as the famous 'blue marble' image of Earth taken by the *Apollo 17* crew had inspired people, Gore said in 1998 that DSCOVR would "awaken a new generation to the environment and educate millions of children around the globe".

To Gore's critics in Congress — particularly Republicans sceptical of his environmental advocacy — those were fighting words. The mission was nothing more than "a multi-million-dollar screen saver", said representative Dave Weldon (Republican, Florida). It acquired the unflattering nickname 'GoreSat'.

"The worst thing that can happen to science is to get mixed up in politics," says Francisco Valero, a retired climate scientist who was at the Scripps Institution of Oceanography in La Jolla, California, and led the satellite's original Earth-science team. "That is what happened to us."

Although Gore intended the project to be mostly educational, NASA formulated a complementary science mission by soliciting proposals from the community. Valero's winning pitch was a probe to measure how much radiation Earth reflects back into space, a crucial variable for untangling the web of processes that influence the planet's climate. Clouds, for example, are a perennial conundrum for climate models because they both reflect incoming sunlight and trap outgoing heat. Valero proposed two instruments: a camera called EPIC to image clouds and other climate-influencing factors such as pollution, volcanic ash and seasonal

KIM SHIPLEY/NASA



vegetation; and a radiometer named NISTAR to measure the energy coming from Earth.

Current estimates of the planet's energy balance rely on stitching together strips of data from orbiting satellites, but DSCOVR will observe the entire sunlit side of Earth. It should reduce errors in estimates of Earth's radiation budget to 1.5% — a more-than-twofold improvement on present measures, says climate scientist Patrick Minnis of NASA's Langley Research Center in Hampton, Virginia, who now leads this aspect of the mission. It will not be the final word, says Adam Szabo, the mission's project scientist at NASA's Goddard Space Flight Center in Greenbelt, Maryland, but DSCOVR will be a big help to climate simulations. And it will partly fulfil Gore's original vision by posting snapshots of Earth online every few hours.

### POLITICAL SCIENCE

Although it blew past its original US\$50-million budget to roughly \$100 million, DSCOVR was still built relatively quickly and cheaply. NASA completed construction in 2000, intending to launch the craft on the space shuttle.

In January 2001, George W. Bush became president after defeating Gore in a controversial election. Soon after, the mission was taken off NASA's shuttle flight manifest. The official reason was that construction of the chronically delayed and over-budget International Space

Station required a higher priority.

But that did not stop speculation about political motives. 'Who killed DSCOVR?' became something of a parlour game in space

***"The worst thing that can happen to science is to get mixed up in politics."***

circles. Mitchell Anderson, a reporter for the climate website DeSmogBlog in Vancouver, Canada, cited an unnamed NASA source who said that Bush's vice-president Dick Cheney had given the order; others suggest that it was the president himself.

In reality, the space shuttle's crowded launch schedule was the biggest obstacle, says Ghassem Asrar, who was the head of NASA's Earth-science division when the decision was made. But the project had become "tainted", he adds, preventing public support from privately sympathetic politicians and from NASA itself. "It would be dishonest to say the politics of climate science wasn't a factor. It was."

In November 2001, with no launch slot in sight, Congress approved \$1 million for DSCOVR to be put into storage at Goddard. And there it might have remained were it not for interest from space-weather forecasters at NOAA and in the Air Force. In 2008, they were looking to cheaply replace NASA's ageing Advanced Composition Explorer, which had

been informing forecasts from the same spot in space that DSCOVR was supposed to occupy. In October that year, Congress ordered NASA to come up with a plan for DSCOVR's revival, and after a series of tests, it began funding NOAA to refurbish and operate the craft with \$105 million over five years.

For Jay Herman, an atmospheric scientist at Goddard and EPIC instrument scientist, the delay has a silver lining: the refurbishment revealed a manufacturing defect in EPIC that would have let in stray light and potentially ruined its image of Earth. The delay allowed enough time to study the problem and correct for it. "So in some ways," says Herman, "I'm very glad it did not fly 14 years ago. Because it might have been embarrassing." ■

### CORRECTION & CLARIFICATION

In the story about Suzanne Topalian in 'Nature's 10' (*Nature* **516**, 311–319; 2014), the text wrongly noted that the July approval for the drug she'd been involved with was in the United States — it was in Japan. The News Feature 'Pollution patrol' (*Nature* **517**, 136–138; 2015) quoted Joshua Apte as saying that air pollution is the largest global health risk. What he meant to say was that it is the largest environmental health risk.





JEFF HUTCHENS/GETTY

Biochar — a soil additive made by heating biological material — is catching attention as a means to improve crop growth and clean up contaminated water.

# STATE-OF-THE-ART SOIL

*A charcoal-rich product called biochar could boost agricultural yields and control pollution. Scientists are putting the trendy substance to the test.*

BY RACHEL CERNANSKY

**F**or more than 150 years, the Brooklyn Navy Yard constructed vessels that helped to stop the slave trade from Africa, lay the first undersea telegraph cable and end the Second World War. Now, this sprawling industrial facility in New York City is filled with artists, architects, producers of artisanal moonshine and people growing organic vegetables. On a drizzly day in autumn, Ben Flanner tends a sea of red and green lettuce on a 6,000-square-metre rooftop farm.

The soil beneath the plants looks ordinary, but Flanner grabs a handful and holds it up for inspection. Amid the brown clods of dirt are small black particles — remnants of charcoal fragments that were mixed into the soil two years ago. Flanner thinks that this carbon-rich material, known as biochar, has helped the crops to thrive, possibly even

increasing their yield, and he hopes for more impressive results over the next few years.

Across the United States, sales of this long-lasting soil additive have surged over the past few years, tripling annually since 2008, according to some estimates. The Biochar Company in Berwyn, Pennsylvania — which supplied Flanner's Brooklyn farm — sells it both wholesale and direct to consumers, through outlets including Amazon and some Whole Foods stores. And countries ranging from China to Sweden are using biochar on agricultural fields and city lawns.

Proponents see big potential for the soil enhancer, which is produced by heating biological material — such as husks and other agricultural waste — in a low-oxygen chamber. Biochar can be made as a by-product of biofuel generation, so some companies are hoping to cash in on both products

as demand grows for greener forms of energy.

Interest in biochar is also growing among scientists, who are quickly ramping up studies to test its potential. They are particularly interested in how the chemical and physical properties of biochar particles affect water moving through soil, remove pollutants, alter microbial communities and reduce emissions of greenhouse gases. The hope is that biochar can help farmers around the world, particularly those in Africa and other developing regions, who often struggle with poor soils.

Johannes Lehmann, a crop and soil scientist at Cornell University in Ithaca, New York, says that different types of biochar “have unique potential to mitigate some of the greatest soil-health constraints to crop productivity — for example, in highly weathered and sandy soils”.

But there are still many questions about biochar, particularly in terms of making sure that it is affordable and has positive effects. In some studies, the material has actually reduced yields. Part of the difficulty is that biochar can be produced from all kinds of biomass and at different temperatures and speeds, which leads to huge variation in the substance — and in results. “I always say we should not even use the singular for biochar,” says Lehmann. “There are only biochars.”

### AMAZONIAN ROOTS

Although it is just starting to catch on with farmers today, biochar has ancient roots. Hundreds to thousands of years ago, residents of the Amazon produced it by heating up organic matter to create rich, fertile soils called *terra preta*. But the practice was abandoned around the time that European nations invaded South America, and relatively few farmers elsewhere have routinely used biochar. Scientists first took a big interest in the material about a decade ago, when growing concerns over global warming led some to tout biochar as a way to store huge amounts of carbon underground. Hope for that application has faded somewhat, but soil scientists are now exploring its use in agriculture and remediating pollution.

A particular focus has been explaining how biochar affects water movement through soils. Rebecca Barnes, a biogeochemist at Colorado College in Colorado Springs, and some of her colleagues tested that by adding biochar to different materials<sup>1</sup>. In sand, through which water typically drains very quickly, biochar slowed the movement of moisture by an average of 92%. In clay-rich soil, which usually retains water, biochar sped up movement by more than 300%.

The researchers suggest that the biochar alters how water moves through the interstitial space — the gaps between grains in the soil.

“Clays tend to be flat grains and sand tends to be circular grains, but biochar is very amorphous — and so it’s not only creating these crazy pathways through the biochar, but it’s also creating crazy pathways in that interstitial space,” says Barnes. She and her colleagues suggest that

these convoluted pathways help to slow down drainage in sand and speed it up in clays.

That is significant, Barnes says, because even though clays can hold large amounts of water, that moisture has a hard time moving through the grains and reaching plant roots. Some studies have shown that plants grow better in soil

## THE HOPE IS THAT BIOCHAR COULD HELP FARMERS, PARTICULARLY IN AFRICA AND OTHER DEVELOPING REGIONS.

with added biochar than in plain soils or those treated just with compost<sup>2</sup>.

Researchers are also teasing apart how biochars influence microbial activity in soil. Microbes typically act as a community; for example, many pathogenic bacteria attack a plant’s roots only when they have sufficient numbers to overwhelm the host’s immune response. Caroline Masiello, a biogeochemist at Rice University in Houston, Texas, and her co-workers have found<sup>3</sup> that biochar can inhibit this by binding to the signalling molecules that bacterial cells secrete to coordinate their activity.

“They all think they’re alone, because the telephone wires have been cut,” says Masiello. With further research, she says, it might be possible to fine-tune this function of biochar to reduce plant infections.

Other researchers are exploring how biochars can cut emissions of nitrous oxide, a greenhouse gas, from agricultural fields. Last year, Xiaoyu Liu, a soil scientist at Nanjing Agricultural University in China, and his colleagues reported<sup>4</sup> that after biochar had been applied to maize (corn) and wheat fields once, nitrous oxide emissions declined over the following five crop seasons, a period of three years. Other studies have shown reductions as well, but researchers have not yet been able to determine what exactly causes this effect. Applying biochar “can also improve some soil properties, like it can increase the potassium availability, and the soil organic-matter content,” says Liu, who has obtained some funding from biochar producers.

But not all studies show biochar to be a wonder material. In some cases it has reduced crop yields<sup>5</sup>, and one study<sup>6</sup> suggests that it lowers the activity of plant genes that help to defend against insect and pathogen attacks.

Lehmann says that this may come down to

improper applications of biochar. In some of the studies that showed decreases in yields, he says, the soils were perfectly fine to start with. Other work suggests<sup>7</sup> that using the wrong type of biochar can negatively impact the soil’s microbiota or, potentially, its carbon-storage capacity. A biochar made from rice straw, for example, will function differently in a certain soil than will biochar made from wood or manure.

Overall, however, the positive impacts of biochar seem to outweigh the negative ones. A 2011 meta-analysis<sup>8</sup> found an overall average yield increase of 10%, rising to 14% in acidic soils. Biochar’s greatest potential might be in places where soils are degraded and fertilizer scarce, in part because it helps the soil to better retain any nutrients that it does have. Andrew Crane-Droesch at the University of California, Berkeley, has been studying the impacts of biochar in such degraded soils in western Kenya. His preliminary data suggest that farms using biochar averaged 32% higher yields than controls.

In June, a World Bank report<sup>9</sup> said that biochar probably holds the most potential for small farmers in developing countries, not just because they are working with the soils most likely to benefit, but because biochar may be a key element of ‘climate-smart’ agriculture — practices that both help to mitigate climate change and reduce vulnerability to its effects.

### POLLUTION WRANGLER

Biochar’s start may have been in agriculture, but researchers are now looking at other applications. Biochar can bind to heavy metals in soil, which helps to keep them from reaching plants or entering water supplies. That has attracted the notice of the US Environmental Protection Agency, other agencies, and companies seeking to reclaim land formerly used in mining. At the Hope Mine near Aspen, Colorado, biochar added in 2010 helped to neutralize the impacts of decades-old mine refuse by immobilizing the metals and increasing the amount of water held on the slope — thereby reducing the opportunity for contaminated water to become run-off. It also helped to spur plant growth on the formerly barren hillside, according to the Aspen Center for Environmental Studies.

Biochar is also showing promise in cleaning up polluted water, perhaps as a much cheaper replacement for activated charcoal, which is used at sites ranging from treatment plants to areas that are heavily contaminated with toxic chemicals. Biochar particles have a relatively large surface area, which expands even further in water, providing a vast number of sites for contaminants to bind to, says Charles Pittman, a retired chemist at Mississippi State University in Starkville. He says that this type of pollution remediation may be particularly beneficial in countries that lack full water-treatment systems. It could also help to remove antibiotics or chemical wastes, which are difficult to strip out with conventional water treatments.

Scientists have even explored biochar’s





ENRIQUE CASTRO-MENDIVIL/REUTERS/CORBIS

Workers at the Villa Carmen Biological Station in Peru turn soil containing black flecks of biochar, produced by burning bamboo in metal drums.

potential for treating fluids used in oil and gas drilling, and as a component of print toners and paint products. “There’s a lot of other markets that haven’t fully been explored yet,” says Kurt Spokas, a soil scientist with the US Department of Agriculture’s Agricultural Research Service in St Paul, Minnesota.

Experts caution, however, that it is not clear when or whether remediation — or other applications — will be economical, particularly in agriculture. Poor soils and poverty often go together. After demonstrating yield increases in Kenya, Crane-Droesch looked at the economic viability of biochar in the same communities. “What we found was almost nobody was willing to pay for biochar when offered at roughly the price it took to make it,” he says.

Biochar prices vary widely, but in the United States some products cost US\$3 per kilogram, comparable to certain fertilizers and more than many composts. On a large scale, biochar production may make economic sense only when biofuel production does — for example if it is subsidized or because policies to reduce carbon emissions drive fossil-fuel prices up.

And if demand ever does surge, there will be questions about the environmental impact of producing biochar. One key concern is the choice of feedstock. China is eager to use agricultural waste, such as rice and wheat straw, and some researchers in the United States are even

pushing animal manure, but neither may be the most efficient way to produce it on a massive scale. And using wood could spur deforestation or harmful land-use practices.

“It’s an incredibly important question to ask: what is the sustainability of the feedstock?” says Alfred Gathorne-Hardy, research director of the India Centre for Sustainable Development at the University of Oxford, UK. “This is the kind of debate I don’t think we’re seeing enough about within the biochar world.”

#### GROWTH INDUSTRY

That debate may grow as consumer interest does — something that is slowly happening around the world. Björn Embrén, who is responsible for tree planning and protection in Stockholm, says that the city has been using biochar to boost local vegetation since 2009; he credits it with the city’s healthiest tree growth in recent years. In September, the New York-based charity Bloomberg Philanthropies awarded Stockholm €1 million (US\$1.2 million) to launch a city-wide programme that will turn residential garden waste, and eventually food waste and even sewage, into biochar.

Back in Brooklyn, Flanner continues to monitor his crops. The lettuces and carrot tops glisten under the rain as he steps carefully between rows in his bright yellow rain jacket. He thinks that the biochar will be good

for his soils over the long term because it helps them to retain nutrients and water. “Those are both very important, especially in such a well-drained soil as on a green roof,” he says. “We tend to lose both of those quickly.”

But before he adds more of the black grains to other parts of his farm, he will wait to see how the crops respond over the next few years. Like the scientists studying biochar, he is eager to see whether it will live up to its bright promise or fade like so many other would-be wonder materials. ■

**Rachel Cernansky** is a freelance writer in Denver, Colorado.

1. Barnes, R. T., Gallagher, M. E., Masiello, C. A., Liu, Z. & Dugan, B. *PLoS ONE* **9**, e108340 (2014).
2. Liu, J. et al. *J. Plant Nutr. Soil Sci.* **175**, 698–707 (2012).
3. Masiello, C. A. et al. *Environ. Sci. Technol.* **47**, 11496–11503 (2013).
4. Liu, X. et al. *Agric. Syst.* **129**, 22–29 (2014).
5. Rajkovich, S. et al. *Biol. Fert. Soils* **48**, 271–284 (2012).
6. Viger, M., Hancock, R. D., Miglietta, F. & Taylor, G. *GCB Bioenergy* <http://dx.doi.org/10.1111/gcbb.12182> (2014).
7. Zimmerman, A. R., Gao, B. & Ahn, M.-Y. *Soil Biol. Biochem.* **43**, 1169–1179 (2011).
8. Jeffery, S., Verheijen, F. G. A., van der Velde, M. & Bastos, A. C. *Agric. Ecosyst. Environ.* **144**, 175–187 (2011).
9. Scholz, S. M. et al. *Biochar Systems for Smallholders in Developing Countries* (World Bank, 2014).





# HERE'S LOOKING AT YOU, SQUID

*Margaret McFall-Ngai has dissected the relationship between a beautiful squid and its live-in bacteria — and found lessons for microbiome research on the way.*

BY ED YONG

**T**he aquarium looks empty, but there is something in it. A pair of eyes stick out from the sandy floor, and their owner is easily scooped up into a glass bowl. At first, the creature looks like a hazelnut truffle — small, round and covered in tiny flecks. But with a gentle shake, the flecks of sand fall off to reveal a female Hawaiian bobtail squid (*Euprymna scolopes*), about the size of a thumb. As she jets furiously around the bowl, discs of pigment bloom and fade over her skin like a living pointillist painting.

There are no other animals in the bowl, but the squid is not alone. Its undersides contain a two-chambered light organ that is full of glowing bacteria called *Vibrio fischeri*. In the wild, their luminescence is thought to match the moonlight welling down from above and cancel out the squid's shadow, hiding the animal from predators. From below, the squid is invisible. From above, it is adorable. "They're just so beautiful,"

says Margaret McFall-Ngai, a zoologist at the University of Wisconsin–Madison. “They’re phenomenal lab animals.”

Few things excite McFall-Ngai more than the partnership between the bobtail squid and *V. fischeri* — and that is after studying it for more than 26 years. Over that time, she has shown that this symbiotic relationship is more intimate than anyone had imagined. She has found that the bacterium out-competes other microbes to establish an entirely faithful relationship with one host. It interacts with the squid’s immune system, guides its body clock and shapes its early development by transforming its body.

Some of these discoveries have helped to shape her field. When McFall-Ngai started her career in 1978, microbiologists were focused almost entirely on pathogens and disease. But in the past decade, advances in genetic sequencing have allowed scientists to identify the trillions of microbes in the bodies of humans and other animals, and to show how they support development, digestion and even behaviour. The study of these communities, collectively known as the microbiome, is now one of the hottest areas in biology, and some of the discoveries made by McFall-Ngai have paved the way. “She pioneered work on animal–microbe interactions well before everyone caught up and the microbiome became such a sexy topic,” says Dianne Newman, a geobiologist at the California Institute of Technology in Pasadena.

The microbiome boom is both a blessing and a curse. Attention and funding has focused heavily on projects to sequence microbes en masse, particularly in the human body, and on efforts to understand how they affect health. The squid and its luminous partner risk being eclipsed, at a time when funding is increasingly tight. But even the most prominent microbiome researchers say that they have time for McFall-Ngai and her squid–bacteria symbiosis, because understanding this simple relationship could help to make sense of more complex microbial communities, which are, by their nature, harder to study. “I’d argue that it’s important to take advantage of the lessons emerging from such systems,” says Jeff Gordon at Washington University in St. Louis, one of the leading figures in human microbiome research. “Their importance isn’t diminished.”

The squid may represent the road less travelled — but McFall-Ngai has always been drawn by such paths. “When I first met her, we were both in LA, driving a lot,” recalls her partner, Ned Ruby. “If she was driving from A to B, even if there was one obvious way, she’d try all these routes. Most would be longer. I’d say, ‘Why are we doing this?’ She’d say, ‘You never know when the freeway’s going to be blocked. I want to scout out the ways of going round.’ That’s how she does science. She doesn’t go down the main road and get blocked. She goes down the side roads.”

## LIGHT THE WAY

McFall-Ngai started down her scientific road as a graduate student, when she became fascinated by bioluminescence and started studying ponyfish, which carry a glowing bacterium. She wanted to understand how these partnerships began, but was frustrated because the fish proved impossible to raise in a lab. Then, a colleague said to her, “Hey, have you heard about this squid?” A few embryologists had been studying the creature, which swims in shallow reef flats around Hawaii and emerges at night to forage. But no one had paid attention to the relationship with its bacteria — until 1988, when McFall-Ngai flew out to Hawaii to take a look.

First, she had to learn how to catch the animals; in knee-deep water, she could snag dozens with just torches and nets. She began breeding them in 1989, when she started her own lab at the University of Southern California in Los Angeles. She found that just 8–10 pairs could produce 60,000 juveniles a year. And unlike animals whose symbionts provide essential nutrients, the squid can survive without *V. fischeri*. This meant that McFall-Ngai could raise the partners separately, introduce them, and watch their first dates.

But first, she needed a collaborator — someone who understood the bacterium. “I think I was the third microbiologist she came to and the first who said yes,” says Ruby. The two had met when they were taking courses in Los Angeles. They have been professional partners ever since she started working with the squid, and romantic ones for most of that

time. “I think it’s a real symbiosis the two of them have,” says Nicole Dubilier from the Max Planck Institute for Marine Microbiology in Bremen, Germany.

McFall-Ngai and Ruby embarked on a journey to unpick every aspect of the squid–bacterium symbiosis, at first in separate institutions, then on adjacent floors at the University of Hawaii at Manoa in Honolulu, and finally in adjoining rooms at the University of Wisconsin–Madison. They knew that the squid are colonized by *V. fischeri* within hours of hatching. But how does the bacterium infiltrate the light organ? And why is it the only species to do so, when other ocean bacteria collectively outnumber it 1,000-fold? To find out, McFall-Ngai carefully dissected the light organ, and Ruby loaded the bacteria with fluorescent proteins to track the microbes’ movements.

Some details of the symbiosis are still falling into place. But the pair now know that the relationship begins on the underside of the newborn squid, when mucus-lined fields of beating hairs called cilia create a current that draws bacteria close. Physics then gives way to chemistry. When *V. fischeri* first touches the squid, it changes the expression of scores of squid genes — a finding<sup>1</sup> made in 2013 by former postdoc Natacha Kremer. Some of these genes produce a cocktail of antimicrobial chemicals that create an inhospitable environment for most microbes while leaving *V. fischeri* unharmed. Others release an enzyme that breaks down the squid’s mucus to produce chitobiose, a substance that attracts more of the bacterium. It takes just five *V. fischeri* cells to trigger these changes, and the microbe soon dominates the fields of cilia (see ‘What the squid hid’).

Chitobiose also stimulates the bacteria to start migrating into three blind-ended crypts in the squid’s light organ. Once they reach their destination, they cause the pillar-like cells that line the crypts to become bigger and denser, enveloping the microbes in a tight embrace<sup>2</sup>. The crypts close off, sealing *V. fischeri* inside for the rest of the squid’s 3–10-month life<sup>3</sup>.

In 2004, McFall-Ngai’s team showed that two molecules carried by the bacteria — peptidoglycan and lipopolysaccharide — are responsible for these changes<sup>4</sup>. That was a surprise. At the time, these chemicals were known only in the context of disease — they were described as pathogen-associated molecular patterns, or PAMPs, tell-tale substances that alert animal immune systems to burgeoning infections. McFall-Ngai took the acronym, swapped the pathogenic P for a microbial M, and rebranded them as MAMPs. These molecules, she proposed, can trigger debilitating inflammation but they can also start a friendship: without them, the squid’s light organ never reaches its mature form.

To McFall-Ngai, these results hinted at a broader theme in biology: animals grow up under the influence of their microbes, not just the blueprints encoded in their genomes. “Most of us would say: Isn’t that interesting? Margaret said: That’s interesting ... and microbes play a role in development,” says Angela Douglas, an entomologist and microbiologist at Cornell University in Ithaca, New York. “She doesn’t deal in little ideas.” McFall-Ngai proposed<sup>5</sup> the concept in 1991, and other scientists have confirmed it, finding that the bodies and immune systems of animals ranging from tsetse flies to mammals mature properly only after exposure to bacteria — sometimes in response to the same MAMPs.

Michael Hadfield, a marine biologist at the University of Hawaii, for example, has shown that the larvae of some marine worms metamorphose into adults only when they encounter bacterial molecules<sup>6</sup>. This made sense when he considered that the earliest animals originated in oceans that were swarming with bacteria. “They very likely evolved to ‘use’ those bacteria as a source of cues for developmental change,” he says.

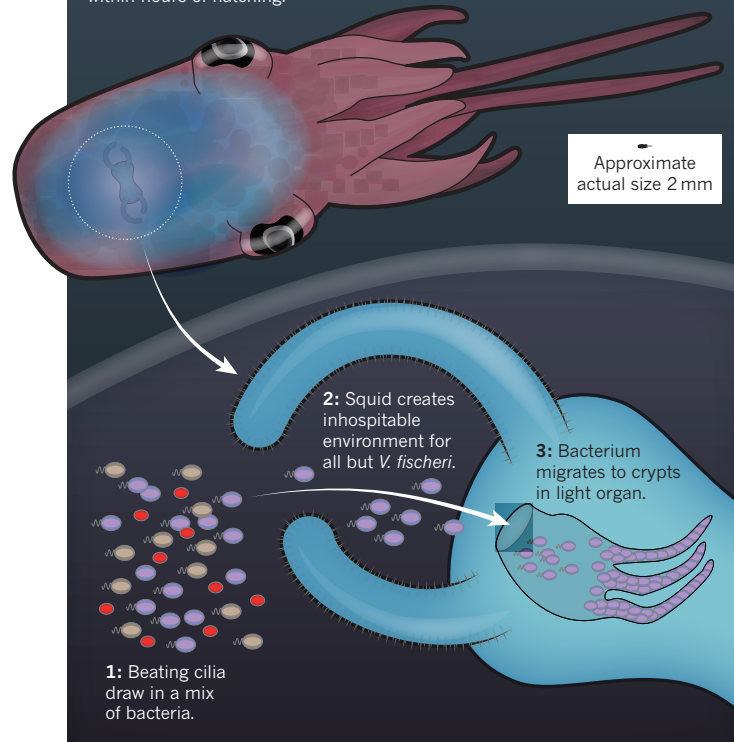
McFall-Ngai has championed other ideas, too. One of them emerged when she started thinking about the adaptive immune system, a trademark of vertebrates that targets incoming microbial threats with bespoke antibodies and retains a memory of past encounters. Invertebrates, including squid, rely on innate immunity — a simpler, short-lived and ever-present battalion of defensive cells. Many immunologists had assumed that vertebrates evolved adaptive immunity because they live longer than invertebrates, and a more complex immune system affords them better protection against pathogens across an extended lifespan.

In 2007 — just as interest in microbiomes was taking off — McFall-Ngai



## WHAT THE SQUID HID

The interaction between *Vibrio fischeri* and the juvenile bobtail squid means that this is the only bacterium to colonize the squid's light organ within hours of hatching.



offered an alternative explanation. In a commentary for *Nature* called *Care for the Community*<sup>7</sup>, she argued that adaptive immunity evolved because vertebrates need to control a more complex microbiome than invertebrates do. They use it to support beneficial microbes and to block those that pose a threat.

Not everyone buys into the hypothesis. Forest Rohwer, an immunologist at San Diego State University in California, points out that corals lack adaptive immunity but host some of the most complicated microbiomes around. Still, he agrees that adaptive immunity might allow vertebrates to fine-tune their large microbiomes, and other scientists concur. “It’s a different way of thinking about the immune system,” says Douglas. “People can agree or disagree with it, but it is a touchstone. If someone says, ‘Remember *Care for the Community*?’, everyone knows what they mean. It’s short for a suite of ideas that challenge traditional notions in a really informed way.”

## TALK ABOUT A REVOLUTION

McFall-Ngai exudes a stateswoman’s confidence as well as a scientist’s exuberance; friends describe her as regal. And so convinced is she by the importance of animal–microbe interactions that her message can verge on evangelism. “We now know that microbes make up the vast diversity of the biosphere, and that animal biology was shaped by interacting with microbes,” she says. “In my mind, this is the most significant revolution in biology since Darwin.”

McFall-Ngai has broadcast that message widely. In 2005, when the American Society for Microbiology was dominated by infectious-disease researchers, she persuaded the organization to run its first meeting on beneficial microbes — a meeting that continues to be popular today. She served on a National Academy of Sciences committee convened by President Barack Obama to outline where biology in the United States will go in the 21st century. In 2012, she helped Newman to create a course that would teach undergraduates the principles of biology using microbes as the starting point of every topic — and she regularly flew from Madison to Pasadena during her holidays to teach the class.

Her passion for the squid has also spawned an academic dynasty. Ruby and McFall-Ngai have now trained dozens of scientists, around 16 of whom are still studying the same symbiosis, now as heads of their own labs. But the duo discourages rivalries. “I grew up watching fields that eat their young, and I didn’t want that,” says McFall-Ngai. The pair invites postdocs who set up their own lab to claim an aspect of the symbiosis for themselves — and every year they host a symposium-party affectionately called the Pow-Wow, at which everyone gets together to share their results and plans. “If someone else says, ‘I was going to do that too’, they sit in a corner and talk about it,” says Ruby.

Despite the conviviality, the group knows that it must compete for a limited pot of funds. “I’ve been told, ‘We’ve already funded Margaret or Ned; how many more can we fund?’” says Spencer Nyholm, an early student of McFall-Ngai who now works at the University of Connecticut in Storrs. “I can’t imagine they would ask this if someone proposed to work with *Drosophila* or *C. elegans* or mice.”

McFall-Ngai says that she and her protégés are just getting started. In one project, she is examining an evolutionary theory predicting that every microbiome should be plagued by cheats — microbes that reap the benefits of life in their hosts but do not provide anything in return. Sure enough, the squid is sometimes colonized by strains of *V. fischeri* that do not make any light. McFall-Ngai’s team has found that the squid can use light-sensitive proteins in its light organ to detect a few dark bacteria among a million brightly glowing ones, and selectively evict them<sup>8</sup>. The team now wants to find out more about how it does this — and the answers might help to explain how humans and other vertebrates manage more complicated microbiomes.

The team has also shown that the squid’s relationship with *V. fischeri* varies over the course of a day, controlling the microbes so that they produce light only at night<sup>9</sup>. And in 2013, former student Elizabeth Heath-Heckman showed that *V. fischeri*, in turn, influences the squid’s daily rhythms through a gene that makes a cryptochrome — a type of protein that affects circadian rhythms in many animals<sup>10</sup>. Cryptochromes are usually activated by environmental light, but Heath-Heckman showed that one of the squid’s cryptochrome genes responds only to the blue light that *V. fischeri* emit, ramping up production of the protein.

On the basis of this work, the team predicted that interactions between people and their resident microbes might also change from day to night — and soon, the evidence was pointing that way. Last year, a group in Israel showed that a significant proportion of microbes in the human gut rise and fall in abundance in a 24-hour cycle<sup>11</sup>, and regular jetlag, for instance, can promote weight gain by disrupting these rhythms.

“One of the things we pound into people who come to the labs is that nobody really gives a damn about the squid,” says Ruby. “They care about the big questions that the squid will help to answer.” To tackle more of those questions, in a few months McFall-Ngai and Ruby will move to share the squid’s home. They will return to Hawaii, where McFall-Ngai will head the Pacific Biosciences Research Center in Honolulu. It is a dream job, and a chance to indulge more in her favourite pastimes — skateboarding and bodysurfing — as well as watch the squid on moonlit nights.

“This was completely backwater science,” she says. “Now it’s front-seat science. It’s been fun to watch people realizing that microbes are the centre of the Universe, and to see the field blossom.” ■

Ed Yong is a science journalist based in London.

1. Kremer, N. *et al. Cell Host Microbe* **14**, 183–194 (2013).
2. Montgomery, M. K. & McFall-Ngai, M. *Development* **120**, 1719–1729 (1994).
3. McFall-Ngai, M. J. & Ruby, E. G. *Science* **254**, 1491–1494 (1991).
4. Koropatnick, T. A. *et al. Science* **306**, 1186–1188 (2004).
5. McFall-Ngai, M. J. & Ruby, E. G. *Science* **254**, 1491–1494 (1991).
6. Hadfield, M. G. *Ann. Rev. Mar. Sci.* **3**, 453–470 (2011).
7. McFall-Ngai, M. *Nature* **445**, 153 (2007).
8. McFall-Ngai, M., Heath-Heckman, E. A., Gillette, A. A., Peyer, S. M. & Harvie, E. A. *Semin. Immunol.* **24**, 3–8 (2012).
9. Boettcher, K. J., Ruby, E. G. & McFall-Ngai, M. *J. Comp. Physiol. A* **179**, 65–73 (1996).
10. Heath-Heckman, E. A. *et al. mBio* **4**, e00167-13 (2013).
11. Thaiss, C. A. *et al. Cell* **159**, 514–529 (2014).



# COMMENT

**POLICY** Collected writings of Bush's science adviser published posthumously **p.268**

**HISTORY** The telling marginalia of alchemical texts **p.269**



**MUSEUMS** A way forward for Italy's natural-history collections **p.271**

**OBITUARY** Mathematician Alexander Grothendieck remembered **p.272**

REUTERS



Flood victims wait to be airlifted from a home near the mouth of the Limpopo River in Mozambique last year.

## Manage climate-induced resettlement

Governments need research and guidelines to help them to move towns and villages threatened by global warming, argue **David López-Carr** and **Jessica Marter-Kenyon**.

**I**nupiaq people are watching in horror as climate change claims their homes. Having endured repeated flooding and erosion from sea-ice melting and permafrost thaw, the 400 residents of Kivalina, an Alaskan village on a low barrier island in the Chukchi Sea, voted in 1998 and in 2000 to relocate — together — to coastal sites on higher ground. More than a decade and a half later, Kivalina remains in limbo, its move stymied by institutional, financial and physical barriers<sup>1</sup>.

No US federal or state agency has a mandate to undertake such mass resettlement,

even though the government spent more than US\$15 million on erosion control between 2006 and 2009. Kivalina has failed to raise funds through climate lawsuits against oil and gas companies. And it has yet to identify suitable relocation sites. Meanwhile, the village's water-supply and waste-storage systems have been damaged, and it could become uninhabitable within a decade.

Tens of thousands of people in more than 85% of Alaska's 213 native villages face similar threats<sup>1</sup>.

► **NATURE.COM**

To hear more about climate resettlement, visit: [go.nature.com/6szkaj](http://go.nature.com/6szkaj)

Papua New Guinea, China and Vietnam have already relocated communities that were vulnerable to flooding<sup>2</sup>. More than a dozen developing countries, including Uganda and Bhutan, have submitted national adaptation plans to the United Nations that involve population resettlement<sup>3</sup>. Sea-level rise this century threatens the cultural and national survival of several low-lying island nations in the Pacific and Indian oceans<sup>4</sup>. By 2050, climate-related hazards such as flooding, soil salinization, coastal erosion and droughts could displace hundreds of millions of people around the world from their homes, either ►

► temporarily or permanently<sup>4</sup>.

In many cases, the best way to protect cultures, livelihoods and social links will be to move as a group. Yet population relocation is practically off the climate-policy radar. The United Nations Framework Convention on Climate Change (UNFCCC) did not officially recognize the need for such resettlement until 2010. And science has barely begun to examine the human and environmental drivers, costs and consequences. How severe must a threat — real or perceived — be for people to feel compelled to move? What determines whether they relocate as individuals or together? And how can the social, economic and psychological downsides of population resettlement be minimized?

Social and environmental scientists and policy-makers need to invest in research to better understand and manage such resettlement. Relocation must be incorporated into climate-adaptation policy discussions and funding initiatives in the run-up to the next UNFCCC Conference of the Parties (COP) in Paris in December.

### LAST RESORT?

In a Thomson Reuters Web of Science search, we identified just 30 papers that examined climate-induced planned relocations. Most concerned coastal communities that were subject to flooding, even though drought is a greater driver of mass displacement. More than half of the case studies were in Alaska or Mozambique.

By contrast, dozens of papers are published each year on migration of individuals and families. But voluntary, managed resettlement of entire towns and villages may be a more effective way to sustain livelihoods and cultures in some cases.

That said, relocating communities is fraught with difficulty. In the past 20 years, more than 300 million people have been

resettled as a result of conservation, urbanization or development schemes, including dam and road-building, mainly in developing countries<sup>5</sup>. Most moves faced local resistance and were detrimental to livelihoods, health and well-being<sup>3,5</sup>. Remuneration for lost income, land and jobs rarely compensated for reduced access to resources, fractured social networks and emotional trauma<sup>2,3</sup>.

Climate-induced resettlement is viewed widely as a last resort<sup>2,4</sup>. Many citizens of the Pacific island of Tuvalu<sup>6</sup> and of coastal northern Australia<sup>7</sup>, for instance, are loathe to relocate in the face of sea-level rise and mounting cyclone intensity. Many members of displaced communities return to their homes once the immediate threat has passed<sup>2,5</sup>. A few months after Cyclone Eline hit Mozambique in 2000, thousands of evacuated flood victims moved back to the Limpopo River valley, pushed by a lack of jobs in the relocation villages and pulled by memories of having coped with earlier floods<sup>8</sup>.

What makes people such as the Kivalinans want to move away for good, and together? Migration becomes almost certain once resilience thresholds are crossed — such as insufficient rain for farmers to grow maize (corn) or the disappearance of an island below rising waters. But even the realization of an impending permanent threat may be enough. By the early 1990s, it had become evident to Kivalinans that their way of life would become unsustainable in their current location within decades.

When trying to decide whether to migrate, sociocultural, political and economic concerns usually trump environmental pressures<sup>5</sup>. The reason most people cite for moving is not to avoid the effects of climate change per se, but to enhance their livelihoods and to remain with family and friends. We expect groups to make judgements along similar lines, while weighing the magnitude of the threat and the resilience of the population.

But cultural and political dynamics could hold even greater sway in decisions related to group resettlement. Cultural systems and traditional livelihoods may be better preserved through relocation than by staying put and defending an existing settlement. Powerful interests, either within or external to the group, can suppress or coerce individual and household mobility.

Place and timing matter. Although some communities will stay until the water laps at their doors, others seem poised to proactively relocate. For example, the Pacific island nation of Kiribati and the city of Miami in Florida are both threatened by sea-level rise. But whereas Kiribati is considering permanent relocation, possibly to Fiji, the people of Miami intend to stay put and are investing hundreds of millions of dollars in shoring up sea walls and drainage systems (see 'On the move'). Such disparities need to be understood if the global community is to manage relocation equitably.

Careful planning is needed. Discussions about resettlement in low-lying island states such as the Maldives may seem premature because sea-level rise is unlikely to inundate these regions for several decades. But with their physical survival and cultural sovereignty at risk, the stakes are high. Planning in advance can save time, money and lives.

### POLICY PRIORITIES

Communication and negotiation are also important. The value that a community places on resettlement compared with other options such as sea-wall construction, temporary emergency evacuation and disaster relief need to be assessed so that appropriate policies can be chosen<sup>2,5</sup>. Coastal cities in the developed world might prefer to invest in flood barriers, for instance. Some vulnerable groups may favour household migration. Temporary measures may be enough when disaster impacts are short-lived or infrequent and emergency relief is available.

Government interventions could fail if officials misinterpret or do not appreciate people's perceptions of risk<sup>5,6,8</sup>. The majority of homes built to rehouse flood-threatened communities elsewhere in Mozambique after Cyclone Eline remained empty because planners underestimated locals' resolve and confidence in dealing with floods, as well as farmers' investment in and reliance on agriculture in the flood plains of their home<sup>8</sup>.

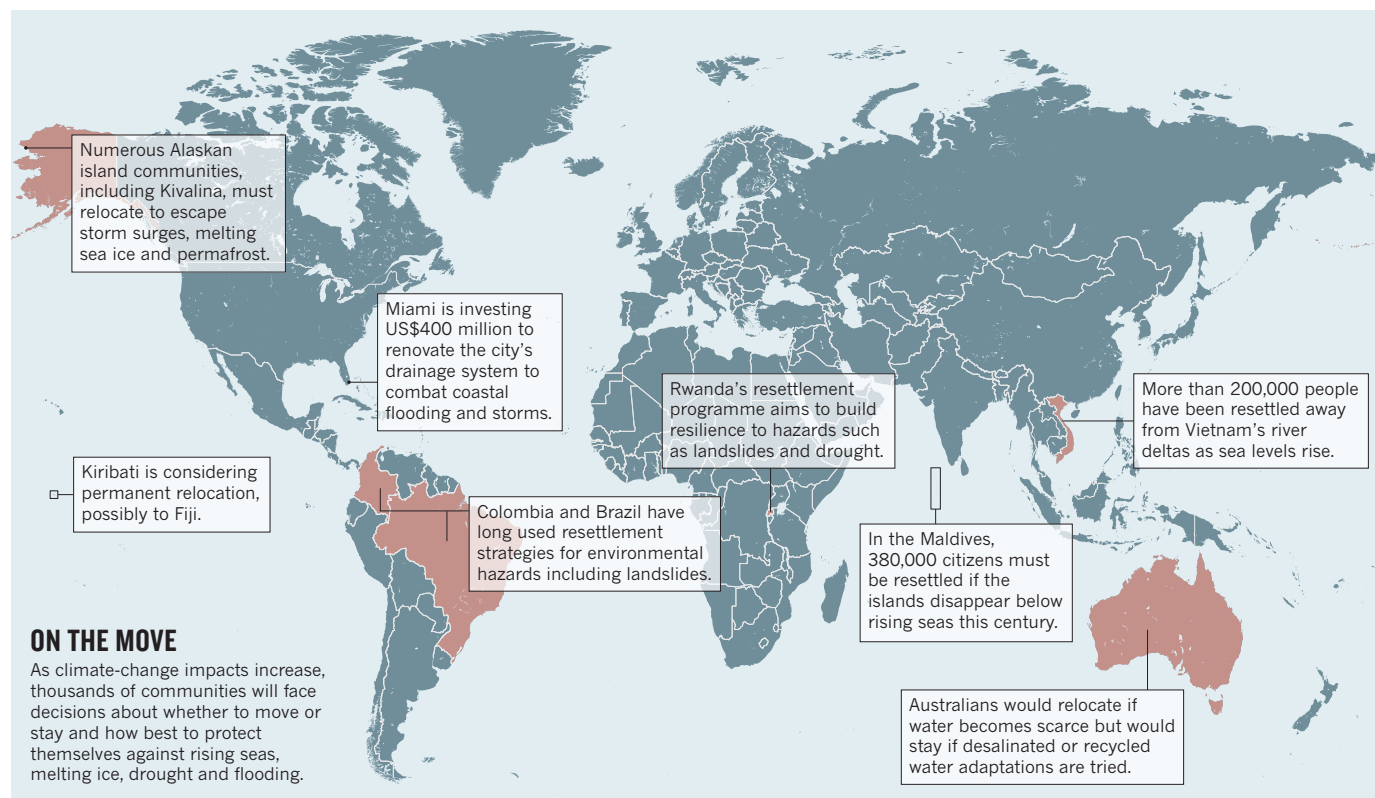
Institutional and legal systems remain ill prepared for managing relocation in response to climate threats. For example, the US Federal Emergency Management Agency, which provides federal aid for preparing for disasters and for relief and recovery after them, has little power to manage pre-emptive resettlement. It can support community relocation only once disaster strikes, and only in response to a handful of hazards — including



When a landslide in Bududa, Uganda, in 2010 wiped out homes and families, the Ugandan government advised people to evacuate.

REUTERS/JAMES AKEVA





drought and hurricanes, but not the severe erosion that threatens Arctic settlements<sup>1</sup>.

Several international policy initiatives have formed recently to develop resettlement guidelines. For example, the Nansen Initiative led by the Norwegian and Swiss governments is planning a global meeting this year to agree to a set of best practices for dealing with cross-border climate displacement. The Peninsula Principles on Climate Displacement within States, developed in 2013 by a group of scientific and legal experts, provides a similar framework for assisting affected people within national borders<sup>2</sup>.

These efforts are a start, but they remain scant and underfunded and are years from application.

### RELOCATION AGENDA

With the worst impacts of climate change yet to come, there is a window of opportunity. Relocation must be higher on the agenda for global climate research and policy; ways to manage it must champion human rights and improve livelihoods. We recommend beginning with the following steps.

**Expand research.** A global survey of climate-induced resettlement and the lessons learned is imperative. Analyses of the drivers, consequences and spatial distributions of resettlement at many scales are crucial to crafting informed policy. The UN, World Bank and Group of 8 nations should convene international working groups on climate relocation. Sustained research funding should be

provided by nations in amounts that are proportionate to their carbon emissions.

**Apply existing international law.** Planning must involve the people who will be affected. Guidelines governing the protection of development- and disaster-displaced people, such as the Hyogo Framework for Action, the UN Principles on Housing and Property Restitution for Refugees and Displaced Persons and the World Bank Guidelines on Involuntary Resettlement, must be applied universally<sup>2,9</sup>. The global community must promote the goals of the Nansen and Peninsula initiatives: to avoid permanent relocation where possible, and to safeguard the needs and rights of people who are displaced by climate change.

**Address constraints to resettlement.** Countries should examine whether their institutional and legal systems are flexible enough to respond to relocation threats — internally and internationally — swiftly, successfully and comprehensively. States will need to adapt laws on disaster declaration and recovery to cover a broader range of environmental hazards, identify funding sources and address climate change within policies on housing, migration and natural-disaster response<sup>3</sup>.

**Improve adaptation financing.** Funding is often an obstacle, even when everyone agrees that relocation is the best option. And resettlement projects are expensive. Billions of dollars will be needed to support the relocation of people in low-income countries.

The UN member states that are responsible for the bulk of carbon emissions should bear the greater burden of this cost.

**Mitigate climate change and minimize its impacts.** Reducing carbon emissions today is more cost-effective and less painful than doing so tomorrow.

In December, the world's climate activists and policy-makers will convene in Paris for the 21st COP for the UNFCCC. There, a global strategy for climate relocation should be developed — for identifying it, understanding it and managing it, equitably. ■

**David López-Carr and Jessica Marter-Kenyon** are in the Department of Geography, University of California, Santa Barbara, Santa Barbara, California 93117, USA.  
e-mail: [jsmarterkenyon@umail.ucsb.edu](mailto:jsmarterkenyon@umail.ucsb.edu)

1. Bronen, R. & Chapin III, F. S. *Proc. Natl Acad. Sci. USA* **110**, 9320–9325 (2013).
2. de Sherbinin, A. et al. *Science* **334**, 456–457 (2011).
3. McDowell, C. *Dev. Policy Rev.* **31**, 677–695 (2013).
4. Barnett, J. & Webber, M. *Accommodating Migration to Promote Adaptation to Climate Change* (World Bank, 2010).
5. Black, R., Arnell, N. W., Adger, W. N., Thomas, D. & Geddes, A. *Environ. Sci. Policy* **27**, S32–S43 (2013).
6. Mortreux, C. & Barnett, J. *Glob. Environ. Chang.* **19**, 105–112 (2009).
7. Zander, K., Petheram, L. & Garnett, S. *Nat. Hazards* **67**, 591–609 (2013).
8. Patt, A. & Schroter, D. *Glob. Environ. Chang.* **18**, 458–467 (2008).
9. *The Peninsula Principles on Climate Displacement Within States* (2013); available at [go.nature.com/8igaqs](http://go.nature.com/8igaqs)





John Marburger was the longest-running head of the US Office of Science and Technology Policy.

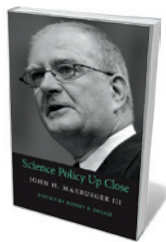
## SCIENCE POLICY

# From Brookhaven to Bush

**Peter Gluckman** finds US presidential science adviser John Marburger's posthumous collection enigmatic.

The complex interfaces between science, society and public policy have evolved considerably over recent decades. The late John Marburger, US President George W. Bush's science adviser from 2001 to 2009, was a transitional and somewhat controversial figure within that time. Across his career, the mood was moving from the unquestioned acceptance of public expenditure on science that had prevailed after the Second World War to greater public interest in the value and societal implications of science — and greater political debate about the role and findings of science.

In his stint as director of Brookhaven National Laboratory in Upton, New York, from 1998 to 2001, Marburger emerged as an insightful intermediary between the scientific community and society. Yet when US science needed those very skills during the Bush era, Marburger seemed strangely silent, at least in public. Nor does this posthumous collection of his writing, *Science*



## Science Policy Up Close

JOHN H. MARBURGER III;  
EDITED BY ROBERT  
P. CREASE  
Harvard University  
Press: 2015.

*Policy Up Close*, reveal his thinking. Its editor, Robert Crease, maintains the silence on this apparent paradox.

But the science-policy nexus has two different faces. Policy advice for science uses levers (such as tax credits for research and development) to influence the national science and innovation ecosystem, most often with economic development as a primary goal. This is distinct from science advice seeking to ensure that public policy is broadly informed by the best available evidence. Generally, both functions are vested in the same office — in the United States, that is the Office of Science and Technology Policy (OSTP). But when Marburger held the combined

posts of presidential science adviser and OSTP director, evidence-informed policy formation was apparently not a priority; and in *Science Policy Up Close*, he does not clarify whether this was by his choice or the administration's directive. Given that the Bush era revealed how vulnerable science can be in the face of organized vested interests such as the oil industry, this omission is frustrating.

Yet Marburger, a physicist, recognized the importance of achieving social licence for technology. This is seen clearly in his work at Brookhaven. In 1997, the lab had discovered a radioactive leak from a storage tank. Marburger realized that it was crucial to involve the local community in discussions. He writes about how — thrown into crisis-management mode over the public's perception of risk at the site — he developed a signature approach to helping stakeholders to see others' views. Similarly, when asking scientists to be accountable for the tax dollars they spend, Marburger was perhaps ahead of his time. His definition of accountability largely excluded scientists' broader contributions to society, but this issue was a theme during his years at the OSTP, and coincided with a governmental turn towards greater investment logic and applying evaluation metrics in managing public science.

In the book, however, Marburger does not address a headline issue of the Bush years — the public shift from trust in science to concerns over the government's treatment of scientific evidence. His only defence is that he chose not to “waste energy” dealing with “controversies that were not in his power to influence”. That he reveals nothing of his own views on issues such as climate change or stem-cell research leaves an uncomfortable vacuum.

Our only window into his thinking is his set of annual addresses to the American Association for the Advancement of Science (AAAS) Forum on Science and Technology Policy, made while he was science adviser and reproduced in the book. In these, Marburger revealed what he felt were the key science-policy issues: scientific-workforce development and scientific immigration in the wake of the terrorist attacks of 11 September 2001; the relative place of funding for discovery science versus directed funding; and approaches to prioritization when national budget appropriations are made.

It was only in 2004 that he seemed to comment on the role of science advice in developing public policy, after the non-profit Union of Concerned Scientists (UCS) issued a statement on ‘Restoring Scientific Integrity to Federal Policy Making’. Marburger's AAAS address that year was a well-scripted riposte to accusations that the Bush administration was ‘anti-science’, stating that

BILL INGALLS/NASA/GETTY

"President Bush believes policy should be made with the best and most complete information". Although 'information' does not necessarily mean 'evidence', that speech remains historically significant as an early instance of the closer relationship between science, society and policy that we know today. (Chris Mooney's 2005 *The Republican War on Science* (Basic Books) offers a very different interpretation of the events surrounding the UCS declaration.)

Marburger's policy comfort zone was clearly the meticulous analysis of the science and innovation ecosystem to better inform the appropriations process. His call for a new "science of science policy" — defining the metrics for evaluating the inputs and outputs of a public science system — is an important legacy that has helped to embed the concept in the government vernacular. That powerful focus on appropriations might have been a strategic way to promote evidence-informed public policy more broadly; but Marburger does not make that claim. It is one thing to support the production of evidence, and quite another to help it to find its way to the corridors of policy. Perhaps Marburger's contribution was in supporting the supply of scientific knowledge, without concerning himself with the more complex business of developing the government's appetite for its systematic use.

*Science Policy Up Close* leaves the impression that Marburger might have had more to say had he been able to finish the book himself. Only the concluding essay offers a hint of his thoughts about the broader role of science in public policy and what he perceived as science's lack of privilege in the seat of US governance.

More than a decade after the UCS declaration, the favoured tactic for dealing with 'inconvenient truths' is perhaps less often about discrediting the science, and more often about acknowledging the evidence and placing it among the many legitimate inputs into policy and decision-making. But there is some way to go: although the science-policy nexus is maturing and becoming more nuanced, the challenges and loneliness of intermediary roles such as Marburger's remain. ■

**Peter Gluckman** is chief science adviser to the prime minister of New Zealand.

## HISTORY OF CHEMISTRY

# Words into gold

**Philip Ball** finds much wrestling with ideas in alchemists' scribbled-over texts.

The sixteenth-century physician and alchemist Paracelsus claimed, "Not even a dog killer can learn his trade from books, but only from experience." As later 'experimental philosophers' turned alchemy into chemistry, they retained this affectation: in the seventeenth century, Robert Boyle is said to have claimed that he had learnt "more from men, real experiments, and his laboratory ... than from books".

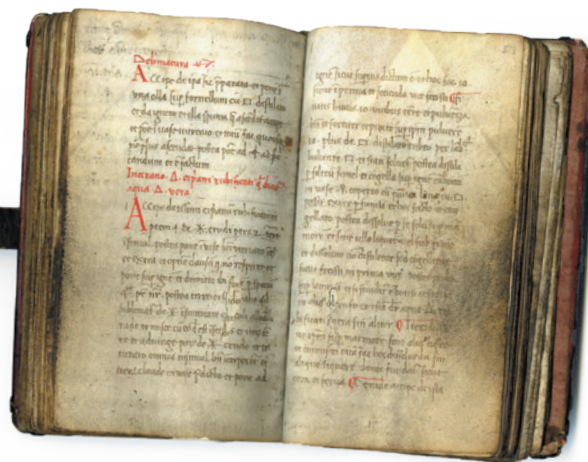
Such comments seem to imply that alchemy and the transitional discipline of 'chymistry' were all about bench-top graft, in contrast to the medieval tradition of seeking knowledge in the library. Yet in most paintings of alchemists at work in the sixteenth and seventeenth centuries, books are ostentatiously on show. Apparatus lies unheeded or broken while the alchemist pores over a text, papers sometimes cascading in comic profusion from desk to floor. In these images, books matter very much indeed: they seem to be where the real secrets lie.

This vexed relationship is examined in *Books of Secrets*, an exhibition at the Chemical Heritage Foundation (CHF) in Philadelphia, Pennsylvania. Juxtaposing fifteenth-century alchemical books and manuscripts recently acquired by the CHF with its extensive collection of paintings of alchemists at their labours, the exhibition explores this early literature of proto-science, and what it was for.

Alchemical books varied significantly. Some were esoteric treatises, all cryptic diagrams and encoded instructions for conducting 'rubification' and other chemical procedures. Others were cheaply printed or hastily copied compilations of miscellaneous recipes for dyes, soaps and medicines. Both were apt to be marketed as 'books of secrets'. The term seems to promise forbidden, mystical insights, but could also simply mean tricks of the trade.

The new acquisitions, originally part of the Bibliotheca Philosophica Hermetica in Amsterdam, include both

**Books of Secrets: Writing and Reading Alchemy**  
Chemical Heritage Foundation, Philadelphia, Pennsylvania.  
Until 4 September.



This alchemical manual may have become soot-smearred over a furnace.

handwritten and printed documents, some attributed (often spuriously) to famous alchemists including Raymond Lull and Petrus Bonus. They reveal the character and functions of the literary culture of nascent chemical science from the Renaissance to the early Enlightenment.

The books were evidently well used. The pages of one fifteenth-century compilation of Italian and English manuscripts arrived covered in dirt — or perhaps soot, from being read over a smoky furnace. The CHF's curator of rare books, James Voelkel, persuaded conservator Rebecca Smyrl to avoid cleaning the pages: the 'dirt' may be a remnant of experiments. "It could be something someone was trying to turn into gold," says Smyrl.

To peruse these books is to glimpse a lively dialogue between author and reader. Despite the volumes' costliness, some have words or passages crossed out or altered. In one sixteenth-century handwritten work, comments are squeezed into every corner of the margins: it is as much lab notebook as reference source.

On this evidence, the painters had it right, even if their depictions of alchemists often owed more to convention than observation. This band of proto-scientists engaged intimately with the words on the page. The text was not sacred, but it was indispensable. ■

**Philip Ball** is a writer based in London.  
e-mail: p.ball@btinternet.com



# Correspondence

## Giant tortoises hatch on Galapagos island

For the first time in 150 years and after more than 50 years of conservation efforts, the number of saddleback giant tortoises (*Chelonoidis ephippium*) seems set to recover unaided on Pinzón Island in the Galapagos archipelago. Rats, early whalers and pirates almost wiped out these ancient creatures.

We found ten tiny, newly hatched saddleback tortoises on the island early last month. There could be many more, because their size and camouflage makes them hard to spot. Our discovery indicates that the giant tortoise is once again able to reproduce on its own in the wild.

The Galapagos National Park and its collaborators set up a programme to save the tortoise in the 1960s, when only about 100 animals remained. This involved collecting eggs and raising hatchlings in captivity for 4–5 years to reach ‘rat-proof’ size, as well as drastic rat-eradication measures (see *Nature* **497**, 306–308; 2013). These strategies have now enabled the species to stabilize itself.

**Washington Tapia Aguilera**  
*Galapagos Conservancy, Santa Cruz, Galapagos, Ecuador.*

**Jeffreys Málaga**  
*Galapagos National Park, San Cristóbal, Galapagos, Ecuador.*

**James P. Gibbs**  
*State University of New York College of Environmental Science and Forestry, Syracuse, USA.*  
[jpgibbs@esf.edu](mailto:jpgibbs@esf.edu)

## Natural history: save Italy's museums

You call attention to the crisis in Italy's natural history museums — in funding, personnel and administration, as well as in their visibility, research and purpose (*Nature* **515**, 311–312; 2014). There is a way to prevent the long-term management of these scientific collections from deteriorating into an elitist hobby. The Italian education and

heritage ministries should together facilitate a functional and administrative connection between the country's smaller natural history museums. Such a ‘meta-museum’ could coordinate long-term goals and scientific activities, enabling facilities and budgetary and technical resources to be shared — as in Germany's Senckenberg Research Institute and Natural History Museum in Frankfurt and Leibniz Association in Berlin. The largest of the museums could stay as independent scientific institutions, similar to London's Natural History Museum and the National Museum of Natural History in Paris.

To strengthen their scientific influence, these museums must participate in survey work and field research that contributes to the discovery, conservation and promulgation of national and global biodiversity.

**Franco Andreone\***  
*Museo Regionale di Scienze Naturali, Turin, Italy.*  
[franco.andreone@regione.piemonte.it](mailto:franco.andreone@regione.piemonte.it)

*\*On behalf of 28 correspondents (see [go.nature.com/c9bcvt](http://go.nature.com/c9bcvt) for full list).*

## Natural history: first museologist's legacy

As the Italian philosopher and naturalist Ulisse Aldrovandi lay ill in November 1603, he dictated his last will and testament — a remarkable and inspiring manifesto of scientific museology. He bequeathed his monumental collections and writings to be held in public trust so that they would be maintained for future generations of scholars. Sadly, his collections are among those now languishing in disarray in Italy (see *Nature* **515**, 311–312; 2014).

Aldrovandi conceived the idea of a natural history museum, the first of which was created as a public institution in Bologna in 1547. He introduced the concept of a sample type for any fossil species, an idea that was

expanded during his lifetime by Francesco Calzolari in Verona, Michele Mercati in Rome and Ferrante Imperato in Naples — underscoring Italy's crucial role in the birth and development of natural history museums.

We have carelessly ignored the will of a great father of museology. It is time to make amends by spurring a renaissance of these museums in Italy.

**Marco Romano**  
*Sapienza University of Rome, Italy.*

**Richard L. Cifelli**  
*Sam Noble Museum, Norman, Oklahoma, USA.*

**Gian Battista Vai**  
*University of Bologna, Italy.*  
[rlc@ou.edu](mailto:rlc@ou.edu)

## Use mentoring to fix science inequality

We suggest that mentorship is particularly important for scientists from the developing world (see *Nature* **515**, 453–454; 2014). It can address the problem of science inequality while helping to resolve global issues.

Academics in developing countries are rarely able to take advantage of cutting-edge knowledge from leading universities, more than 90% of which are in high-income nations (see [go.nature.com/wffbf](http://go.nature.com/wffbf)). Expenditure on research and development is typically less than 1% of gross domestic product — one-third of the amount spent by most developed nations (see [go.nature.com/gio8pu](http://go.nature.com/gio8pu)).

Scientists from wealthy countries could be encouraged to mentor young scientists from developing nations by including mentorship requirements in grant awards, for example.

Also, allocating resettlement funding can help to counter the ‘brain drain’ of researchers leaving low-income countries. It encourages them to return to share experience and knowledge and build local capacity.

**Malgorzata Blicharska**  
*Swedish Biodiversity Centre, Uppsala, Sweden.*

**Grzegorz Mikusiński**  
*Swedish*

*University of Agricultural Sciences, Skinnskatteberg, Sweden.*  
[malgorzata.blicharska@slu.se](mailto:malgorzata.blicharska@slu.se)

## IPBES responds on conflicts of interest

Contrary to the impression given by Axel Hochkirch and colleagues (*Nature* **516**, 170; 2014), the Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES) is expected to approve a policy on conflicts of interest for authors at this week's third plenary in Bonn, Germany.

The IPBES has rules and procedures for nominating and selecting experts from a wide variety of sectors to ensure credibility and transparency (see [go.nature.com/zh1osy](http://go.nature.com/zh1osy)). Accordingly, these experts are nominated by governments and stakeholders and selected by our Multidisciplinary Expert Panel.

Our interim conflict-of-interest policy is in operation for the pollinators assessment (completed declaration forms available from the secretariat on request). The scientists from agrochemical companies mentioned by Hochkirch *et al.* were selected in their capacity as independent scientists to provide objective input for the pollination report. Moreover, independent peer review of all IPBES texts guards against any bias.

Reports by other global and regional assessment panels routinely include authors from commercial sectors — for example, the Intergovernmental Panel on Climate Change and the International Assessment of Agricultural Knowledge, Science and Technology for Development. The latter was even accused of being unbalanced when industrial scientists withdrew from the assessment process (see *Nature* **451**, 223–224; 2008).

**Anne Larigauderie\***  
*IPBES, Bonn, Germany.*

[anne.larigauderie@ipbes.net](mailto:anne.larigauderie@ipbes.net)  
*\*On behalf of 4 correspondents (see [go.nature.com/swyese](http://go.nature.com/swyese) for full list).*



# Alexander Grothendieck

## (1928–2014)

Mathematician who rebuilt algebraic geometry.

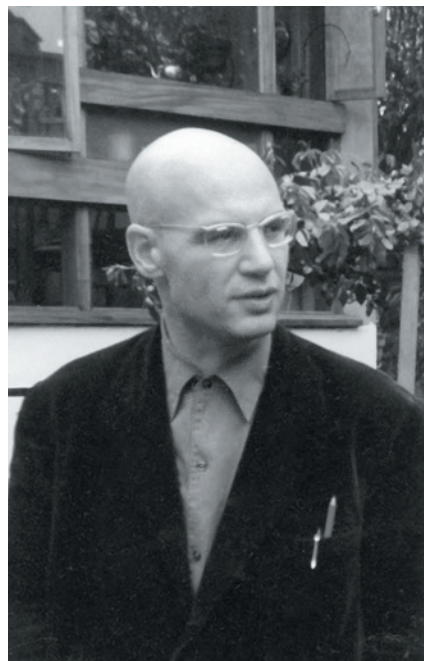
Alexander Grothendieck, who died on 13 November, was considered by many to be the greatest mathematician of the twentieth century. His unique skill was to burrow into an area so deeply that its inner patterns on the most abstract level revealed themselves, and solutions to old problems fell out in straightforward ways.

Grothendieck was born in Berlin in 1928 to a Russian Jewish father and a German Protestant mother. After being separated from his parents at the age of five, he was briefly reunited with them in France just before his father was interned and then transported to Auschwitz, where he died. Around 1942, Grothendieck arrived in the village of Le Chambon-sur-Lignon, a centre of resistance against the Nazis, where thousands of refugees were hidden. It was probably here, at the secondary school Collège Cévenol, that his fascination for mathematics began.

In 1945, Grothendieck enrolled at the University of Montpellier. He completed his doctoral thesis on topological vector spaces at the University of Nancy in 1953, and spent a short time teaching in Brazil. His most revolutionary work happened between 1954 and 1970, mainly at the Institute of Advanced Scientific Studies (IHÉS) in a suburb of Paris. His strength and dedication were legendary: throughout his 15 years in mainstream mathematics, he would work long hours in the unheated attic of his house seven days a week. He was awarded the Fields medal in 1966 for his work in algebraic geometry.

Algebraic geometry is the field that studies the solutions of sets of polynomial equations by looking at their geometric properties. For instance, a circle is the set of solutions of  $x^2 + y^2 = 1$ , and in general such a set of points is called a variety. Traditionally, algebraic geometry was limited to polynomials with real or complex coefficients, but just before Grothendieck's work, André Weil and Oscar Zariski had realized that it could be connected to number theory if you allowed the polynomials to have coefficients in a finite field. These are a type of number that are added like the hours on a clock — 7 hours after 9 o'clock is not 16 o'clock, but 4 o'clock — and it creates a new discrete type of variety, one variant for each prime number  $p$ .

But the proper foundations of this enlarged view were unclear, and this is where, inspired by the ideas of the French mathematician Jean-Pierre Serre, but generalizing them



enormously, Grothendieck made his first hugely significant innovation. He proposed that a geometric object called a scheme was associated to any commutative ring — that is, a set in which addition and multiplication are defined and multiplication is commutative,  $a \times b = b \times a$ . Before Grothendieck, mathematicians considered only the case in which the ring is the set of functions on the variety that are expressible as polynomials in the coordinates. In any geometry, local parts are glued together in some fashion to create global objects, and this worked for schemes too.

An example might help in illustrating how novel this idea was. A simple ring can be generated if we make a ring from expressions  $a + b\varepsilon$ , in which  $a$  and  $b$  are ordinary real numbers but  $\varepsilon$  is a variable with only 'very small' values, so small that we decide to set  $\varepsilon^2 = 0$ . The scheme corresponding to this ring consists of only one point, and that point is allowed to move the infinitesimal distance  $\varepsilon$  but no further. The possibility of manipulating infinitesimals was one great success of schemes. But Grothendieck's ideas also had important implications in number theory. The ring of all integers, for example, defines a scheme that connects finite fields to real numbers, a bridge between the discrete and classical worlds, having one point for each prime number and one for the classical world.

Probably his best-known work was

his discovery of how all schemes have a topology. Topology had been thought to belong exclusively to real objects, such as spheres and other surfaces in space. But Grothendieck found not one but two ways to endow all schemes, even the discrete ones, with a topology, and especially with the fundamental invariant called cohomology. With a brilliant group of collaborators, he gained deep insight into theories of cohomology, and established them as some of the most important tools in modern mathematics. Owing to the many connections that schemes turned out to have to various mathematical disciplines, from algebraic geometry to number theory to topology, there can be no doubt that Grothendieck's work recast the foundations of large parts of twenty-first-century mathematics.

Grothendieck left the IHÉS in 1970 for reasons not entirely clear to anyone. He turned from maths to the problems of environmental protection, founding the activist group *Survivre*. With a breathtakingly naive spirit (that had served him well in mathematics), he believed that this movement could change the world. When he saw that it was not succeeding, in 1973, he returned to maths, teaching at the University of Montpellier. Despite writing thousand-page treatises on yet-deeper structures connecting algebra and geometry (still unpublished), his research was only meagrely funded by the CNRS, France's main basic-research agency.

Grothendieck could be very warm. Yet the nightmares of his childhood had made him a complex person. He remained on a Nansen passport his whole life — a document issued for stateless people and refugees who could not obtain travel documents from a national authority. For the last two decades of his life he broke off from the maths community, his wife, a later partner and even his children. He sought total solitude in the village of Lasserre in the foothills of the Pyrenees. Here he wrote remarkable self-analytical works on topics ranging from maths and philosophy to religion. ■

**David Mumford** is professor emeritus at Harvard University in Cambridge, Massachusetts, and at Brown University in Providence, Rhode Island, USA. **John Tate** is professor emeritus at Harvard University, and at the University of Texas at Austin. e-mails: dbmumford@gmail.com; tate@math.utexas.edu

H. VAN REGEMORTER/IHÉS

## Boxed up and ready to go

**Flow-tank experiments and fluid-dynamics simulations refute the idea that water movements over the body of boxfishes are a stabilizing influence, instead showing that the fish's shape amplifies destabilizing forces to improve manoeuvrability.**

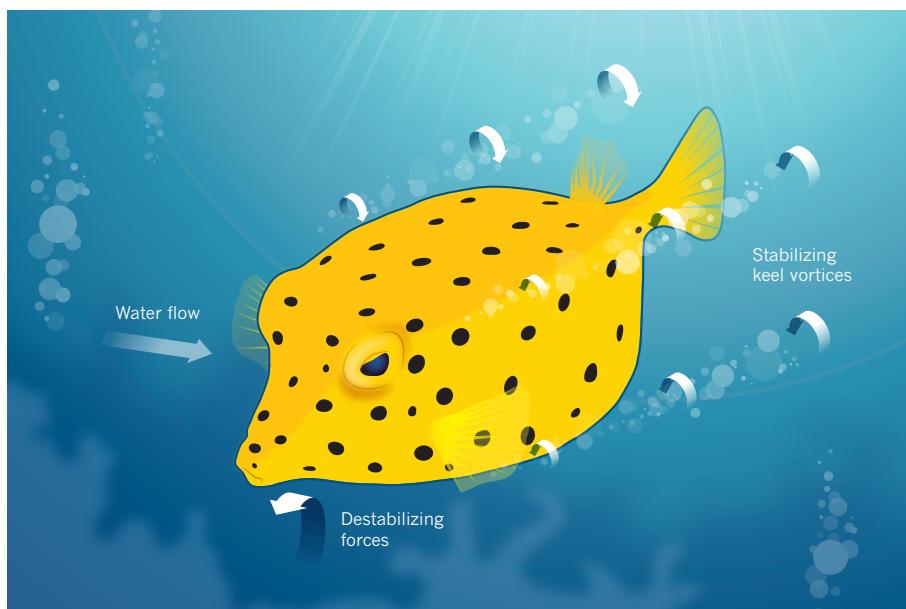
STACY C. FARINA & ADAM P. SUMMERS

A curious denizen of reefs, the boxfish flits among the corals, turning this way and that as it feeds on small invertebrates. The fish has been used as the biomimetic model for a low-drag concept car, the Mercedes-Benz Bionic, but it does not seem to be hydrodynamically gifted, because its swimming is neither swift nor effortless. Writing in the *Journal of the Royal Society Interface*, Van Wassenbergh *et al.*<sup>1</sup> measured the hydrodynamic properties of the boxfish shape by using three-dimensional (3D) printed models and computer simulations of fluid flow. They found that, contrary to previous research, the shape of the boxfish's body creates high drag and passively lends itself to destabilizing flow.

The boxfish would be an ideal costume for a fancy-dress party, because it can be modelled with a painted cardboard container with holes for the head, arms and legs. Its stiff carapace is an external skeleton made of plate-like, fused scales with large keels, like the edges of a box, that run along the length of the body (Fig. 1). Most other fishes power their swimming by moving their muscular bodies and tail from side to side, whereas the inflexible boxfish waggles its pectoral and pelvic fins, aided by occasional steering from the tail. Its progress through the water is largely determined by the shape of its wrap-around armour.

For swimming and flying animals, stability and manoeuvrability are opposing needs, with a gain in stability usually meaning a loss of manoeuvrability. Observations of swimming boxfishes have shown that they are highly stable during straightforward swimming, only infrequently being pushed off course by ambient flows and their own body movements<sup>2,3</sup>. Data from models of boxfishes suggest drag coefficients less than one-fifth of that of a cube moving through the water. These observations led to the boxfish being touted as a model for a low-drag, high-stability shape for a high-volume structure<sup>3</sup>. However, boxfishes are also extremely manoeuvrable, able to make 180° turns in the length of their body<sup>4</sup>. This presents a paradox — how can the carapace of a boxfish provide stability without inhibiting manoeuvrability?

A series of previous studies had suggested



**Figure 1 | Boxfish instability.** The external skeleton of boxfishes — the carapace — is made up of rigid, fused scales. The edges of this carapace are called keels. Previous research<sup>5–7</sup> had suggested that water flow leads to vortices forming around the keels that stabilize the boxfishes' movements. However, Van Wassenbergh *et al.*<sup>1</sup> now show that the effect of these stabilizing vortices is outweighed by the destabilizing forces generated by the boxy front of the boxfish carapace, and this overall instability is what gives the boxfish its remarkable manoeuvrability.

that the boxfish carapace is self-stabilizing<sup>5–7</sup>. The authors of these studies proposed that, when the fish is thrown off course by turbulence, vortices form around the keels, pushing the fish back into a forward-facing position, so the keels act like the stabilizing flights of a dart. This passive process would require no energy or neural input from the fish, allowing instantaneous and inexpensive stabilization. Using 3D models of carapaces in a flow tank, the researchers visualized the vortices responsible for this self-stabilization and consistently found vortices occurring in positions that would provide stabilizing forces.

However, in the latest study, Van Wassenbergh and colleagues quantified flow around the entire carapace, not just around the keels, and found that the overall shape of the boxfish is actually destabilizing. The authors scanned the surface of the carapaces of two boxfish species to create 3D models, then used computational fluid dynamics to analyse the flows around each digitized shape. Drag

measurements are fraught with pitfalls, but these physical and computational models put the boxfish drag coefficient at twice the previous values, which seems more likely. But the really interesting results addressed the paradox of a stabilized yet manoeuvrable fish.

On the basis of the shape of the carapace alone, a boxfish thrown off course by the current, or by its own fin movements, would tend to continue turning in the direction in which it had been pushed. Van Wassenbergh and colleagues visualized vortices trailing from the keels of the carapace and also observed that these vortices produce stabilizing forces — just as the previous studies had shown. However, these forces were not nearly strong enough to overcome the larger destabilizing forces at the front of the carapace (Fig. 1). These destabilizing forces lead to great manoeuvrability, so the boxfish gains passive amplification of its movements from its shape. The authors corroborated this finding by printing their 3D models and measuring the forces acting on



them in a flow tank under a variety of flow conditions. They again found that the shape of the carapace amplified the movements of the boxfishes, rather than stabilizing them.

So it seems that, far from being darts gliding across the reef in a stable manner, boxfishes are tumblers, able to exploit small asymmetrical force inputs at the front of the carapace to generate large changes in direction. This raises an entirely different possibility for biomimetic applications, because the most manoeuvrable, low-radar-signature fighter jets, such as the F-117 Nighthawk, are also dynamically unstable<sup>8</sup>.

If the boxfish carapace has high drag and

is unstable, how was Mercedes-Benz able to model a low-drag car inspired by its shape? The answer lies in the nose of the car, which is rounded and so does not reflect the boxy front of the boxfishes. The front of the carapace amplifies the upsetting force, whereas the boxfishes' keels are stabilizing. By retaining the keels but omitting the boxy head, the car combines stability with low drag. ■

**Stacy C. Farina** is in the Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA.

**Adam P. Summers** is at Friday Harbor Laboratories, University of Washington,

Friday Harbor, Washington 98250, USA.

e-mails: stacy.farina@gmail.com;

fishguy@uw.edu

1. Van Wassenbergh, S., van Manen, K., Marcroft, T. A., Alfaro, M. E. & Stamhuis, E. J. *J. R. Soc. Interface* **12**, 20141146 (2014).
2. Hove, J. R., O'Bryan, L. M., Gordon, M. S., Webb, P. W. & Weihs, D. J. *Exp. Biol.* **204**, 1459–1471 (2001).
3. Bartol, I. K., Gordon, M. S., Webb, P. W., Weihs, D. & Gharib, M. *Bioinsp. Biomim.* **3**, 014001 (2008).
4. Walker, J. A. *J. Exp. Biol.* **203**, 3391–3396 (2000).
5. Bartol, I. K. et al. *Integr. Comp. Biol.* **42**, 971–980 (2002).
6. Bartol, I. K. et al. *J. Exp. Biol.* **206**, 725–744 (2003).
7. Bartol, I. K., Gharib, M., Webb, P. W., Weihs, D. & Gordon, M. S. *J. Exp. Biol.* **208**, 327–344 (2005).
8. Crickmore, P. F. & Crickmore, A. J. *Nighthawk F-117 Stealth Fighter* (Zenith, 2003).

## EARTH SCIENCE

# Mixing it up in the mantle

**Analysis reveals that the uranium isotopic composition of oceanic crust that is being subducted into Earth's interior is distinctive, allowing the development of chemical heterogeneity in the mantle to be tracked. SEE LETTER P.356**

JON WOODHEAD

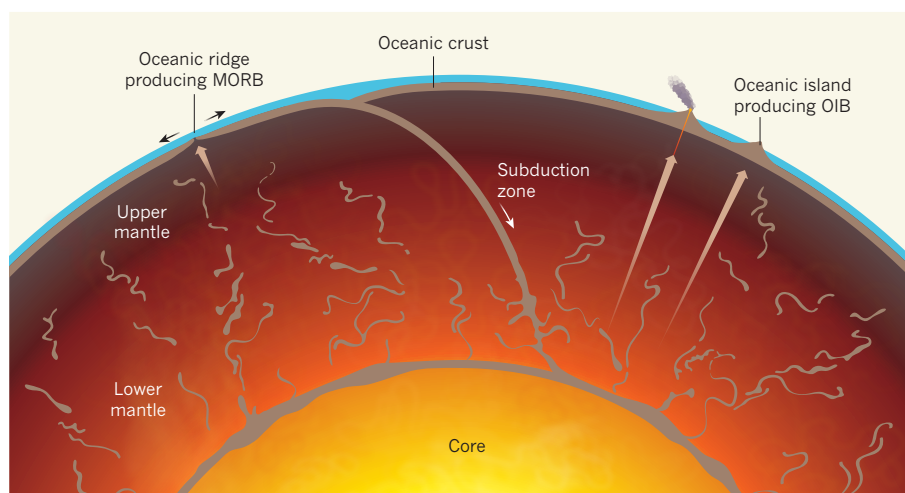
It is now more than three decades since researchers first proposed<sup>1</sup> the radical hypothesis that oceanic crust, returned to the mantle (or subducted) during collisions between tectonic plates, could strongly influence the chemistry of Earth's interior, and furthermore, that the tell-tale signatures of this process could be seen in the volcanic products of mantle melting. In particular, the chemical traces of such 'crustal recycling' (Fig. 1) phenomena could be discerned in rocks termed ocean island basalts (OIBs) that are associated with volcanic 'hotspots' such as Hawaii. Variations on this simple yet provocative idea have provided a focal point for studies of mantle geochemistry and planetary evolution ever since. However, despite a substantial research effort and considerable advances in our understanding, definitive estimates of the timing of crustal-material transport into the mantle have remained elusive. In this issue, Andersen *et al.*<sup>2</sup> (page 356) report on how a relatively new approach, using isotope ratios of the element uranium, provides some long-awaited temporal constraints on these crustal-recycling processes.

It is generally accepted that recycled materials have a key role in the generation of compositional heterogeneity in Earth's mantle<sup>3</sup>, and indeed, evidence to this effect continues to appear<sup>4</sup>. By contrast, the question of when the mantle became modified in this way has proved remarkably intractable. For example,

the abundances of the isotopes of lead (Pb) — derived from slow decay of long-lived uranium (U) and thorium (Th) parent nuclei — in OIBs form linear correlations that suggest a broad, model-dependent age range (about 2.5 billion to 1 billion years) for the establishment of isotopic heterogeneity in their mantle source<sup>5</sup>. Another temporal constraint is provided by the unusually low abundance ratios

of thorium to uranium (Th/U) observed in basaltic lavas erupted at Earth's ocean ridges, known as mid-ocean-ridge basalts (MORBs). These ratios are lower than those estimated for the bulk Earth and have been explained<sup>6</sup> as resulting from uranium recycling into the mantle at subduction zones, perhaps starting about 2.4 billion years ago, coincident with the rise of atmospheric oxygen (and hence the availability of water-soluble hexavalent uranium, U(VI)). Beyond these few, rather imprecise estimates, we have scant information on the timescales of crustal-recycling phenomena.

Isotope geochemistry has always been an instrument-intensive discipline, quick to embrace new opportunities provided by technological advances. The introduction of mass spectrometers known as multiple-collector inductively coupled plasma mass spectrometers (MCICPMS) over the past 20 years has allowed detailed investigations of isotopic systems previously beyond our analytical capability, resulting in many breakthroughs.



**Figure 1 | Crustal recycling.** Oceanic crust (brown) is 'recycled' into Earth's mantle at convergent plate boundaries (subduction zones). Over time, this crustal-recycling process has formed a chemically heterogeneous mantle mixture. Andersen and colleagues' results<sup>2</sup> place constraints on the timing of these events. They suggest that the upper-mantle source producing mid-ocean-ridge basalts (MORBs; short brown arrow) was contaminated in this way over the past 0.6 billion years, whereas heterogeneity in the deeper-mantle source producing ocean island basalts (OIBs; long brown arrows) probably resulted from a much older period of contamination between 0.6 billion and 2.5 billion years ago.

One of these has been the discovery that the abundance ratio of the uranium isotopes  $^{238}\text{U}$  to  $^{235}\text{U}$  ( $^{238}\text{U}/^{235}\text{U}$ ), long held to be invariant in nature, shows small variations<sup>7,8</sup>. In their study, Andersen *et al.* used MCICPMS instrumentation to explore the ramifications of this new paradigm for Earth's uranium-isotope cycle. They detected subtle variations in  $^{238}\text{U}/^{235}\text{U}$  in a range of geological samples, including OIBs and MORBs, and demonstrated that the present-day oceanic crust being subducted into the mantle is isotopically distinct from the bulk Earth, with high  $^{238}\text{U}/^{235}\text{U}$  values. Furthermore, they showed that this feature probably results from the emergence of fully oxygenated oceans 0.6 billion years ago.

Importantly, the authors then observed that the mantle sources of MORBs and OIBs responded differently to subduction of this isotopically distinct crust. Whereas the shallower MORB source also shows high  $^{238}\text{U}/^{235}\text{U}$  values, suggesting widespread pollution of its mantle source by subducted crust predominantly in the past 0.6 billion years, the deeper OIB reservoir shows no sign of this effect. Andersen and colleagues propose that this reflects the greater antiquity of the recycling events contributing to the OIB reservoir. When coupled with previous observations of Th/U, these new data suggest ages for the deep OIB reservoirs of between 2.5 billion years (derived from model ages based on the abundance of Pb isotopes) and 0.6 billion years (when the oceans became fully oxygenated).

As exciting as these results are, we must realize that the range of  $^{238}\text{U}/^{235}\text{U}$  variations observed is at the limit of what current instrumentation can detect. In addition, there are currently no appropriate reference materials characterized to this level of precision. For this reason, further detailed studies will be required to confirm these remarkable observations. Moreover, in an attempt to characterize 'average' oceanic crust, the authors used composite samples of oceanic crust, derived from Ocean Drilling Program Site 801 in the western Pacific. It is well known that the oceanic crust is highly heterogeneous and thus a crucial goal for future studies will be to broaden the uranium-isotope database to determine whether the crustal composite from Site 801 is truly representative of subducting oceanic crust.

It is also interesting that recent *in situ* analyses of sulfur isotopes in tiny olivine-hosted sulfide inclusions in OIB lavas from Polynesia<sup>9</sup> have revealed isotopic compositions that could have been generated on the early Earth by photochemical reactions only before about 2.45 billion years ago, providing a lower age limit for the mantle source of these particular lavas. At first sight, these results seem inconsistent with those of Andersen *et al.*, but we need to bear in mind that both studies encompass only a small part of the diverse OIB compositional spectrum. Future investigations exploiting these emerging geochemical tools will need

to examine the many other OIB flavours: only then may we finally unlock the secrets of crustal recycling. ■

**Jon Woodhead** is in the School of Earth Sciences, University of Melbourne, Victoria 3010, Australia.  
e-mail: [jdwood@unimelb.edu.au](mailto:jdwood@unimelb.edu.au)

1. Hofmann, A. W. & White, W. M. *Earth Planet. Sci. Lett.* **57**, 421–436 (1982).

2. Andersen, M. B. *et al. Nature* **517**, 356–359 (2015).
3. Stracke, A., Bizimis, M. & Salters, V. J. M. *Geochim. Geophys. Geosys.* **4**, 8003 (2003).
4. Workman, R. K., Hart, S. R., Eiler, J. M. & Jackson, M. G. *Geology* **36**, 551–554 (2008).
5. Chase, C. G. *Earth Planet. Sci. Lett.* **52**, 277–284 (1981).
6. Elliot, T., Zindler, A. & Bourdon, B. *Earth Planet. Sci. Lett.* **169**, 129–145 (1999).
7. Stirling, C. H., Andersen, M. B., Potter, E.-K. & Halliday, A. N. *Earth Planet. Sci. Lett.* **264**, 208–225 (2007).
8. Weyer, S. *et al. Geochim. Cosmochim. Acta* **72**, 345–359 (2008).
9. Cabral, R. A. *et al. Nature* **496**, 490–494 (2013).

## GENOMICS

# African dawn

**The African Genome Variation Project presents genotyping and whole-genome data from individuals across sub-Saharan Africa, giving insight into population history and guiding future genomic studies on the continent. [SEE ARTICLE P.327](#)**

**RAJ RAMESAR**

**T**he story of human origins and diversification is the story of Africa, and there is a growing interest in this story being developed and told by Africans. The paper on the African Genome Variation Project by Gurdasani *et al.*<sup>1</sup>, which is published on page 327 of this issue, is an excellent example of collaborative work by African and non-African researchers, and a timely addition to the trickle of human-genomic data aimed at systematically characterizing genetic diversity across Africa. The project presents genotype data from more than 1,481 individuals from 18 ethnolinguistic groups in sub-Saharan Africa (Fig. 1), and whole-genome sequencing from 320 individuals representing 7 ethnolinguistic groups from 3 geographically distinct regions — Ethiopia (northeast Africa), Uganda (east Africa) and southern Africa. The resource is the most comprehensive representation of African diversity so far.

Gurdasani and colleagues' paper uses data from 2,864 individuals from 33 African and non-African populations to provide a bird's-eye view of the dynamicity of the human genome and fine-tuned information on areas of genomic differentiation and fixation in certain African populations. It also highlights the usefulness of this information for identifying genes related to disease susceptibility and resistance. The authors' selection of populations representative of the three major African ethnolinguistic groups (Niger-Congo, Nilo-Saharan and Afro-Asiatic) takes into account the fact that these groupings represent key points — source, nexus and/or destination — on the routes of early human migration, and thus reflect interesting patterns of genetic diversity and selection.

A noteworthy example among the authors'

findings is signatures of significant mixing between Eurasian and west African genomes suggesting that Eurasians migrated back into Africa in the period around 7,500 to 10,500 years ago, following the original exodus of modern humans from Africa that occurred tens of thousands of years earlier. The authors also detect imprints of Khoe-San populations from southern Africa in modern west African genomes. Although others have reported the presence of genetic 'tracks' from Eurasian and Khoe-San populations in the genomes of modern east and southern Africans<sup>2,3</sup>, finding such traces in west African genomes is interesting because it is probable that this Khoe-San

admixture represents an ancient population.

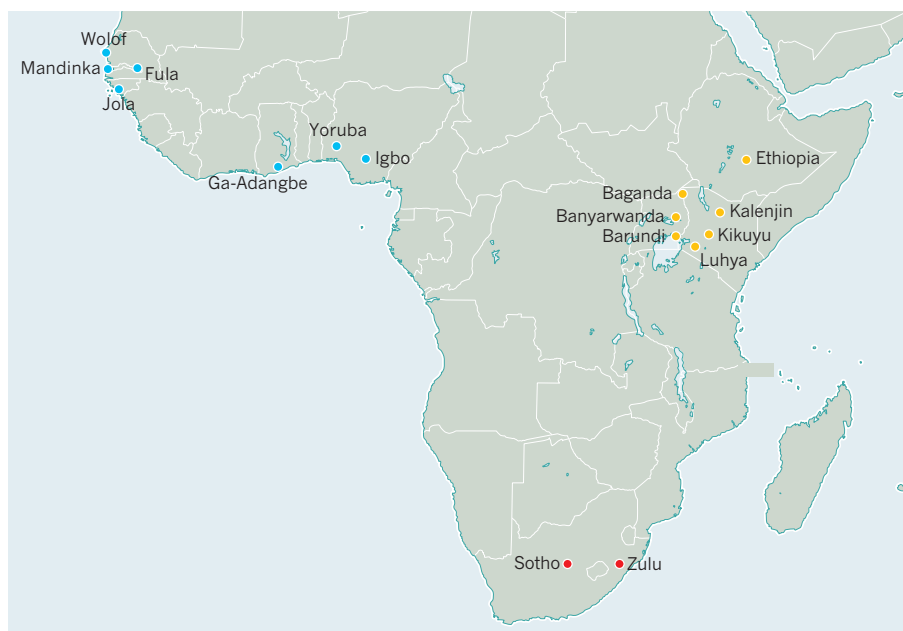
**The three major African ethnolinguistic groups studied represent key points on the routes of early human migration.**

A possible interpretation of these findings is that these Eurasian immigrants, or indeed the Khoe-San, brought with them a 'wanderlust' gene that was then integrated into other

African groups and translated into the Bantu expansion — the series of migrations, occurring around 3,000–5,000 years ago, that spread the Niger-Congo ethnolinguistic group across much of sub-Saharan Africa. Large-scale whole-genome sequencing across Africa will provide many more insights into human diversity, evolution, population history and disease susceptibility, and emerging work on ancient African genomes will further help to resolve unattributed genetic diversity in African populations.

Gurdasani *et al.* also find that the Afro-Asiatic and Nilo-Saharan language groups make a larger contribution to African differentiation





**Figure 1 | African genetic diversity.** The African Genome Variation Project presents<sup>1</sup> genomic information from 18 ethnolinguistic groups across sub-Saharan Africa, representing populations from the west, east and south of the continent ('Ethiopia' encompasses the Oromo, Amhara and Somali ethnolinguistic groups). Locations of spots are approximate.

and diversity than previously estimated, suggesting that the Niger-Congo language groups (which represent the majority of the population across sub-Saharan Africa) were founded fairly recently by large numbers of individuals, coinciding with the Bantu expansion. The authors also observed a remarkable decrease in sub-Saharan population differentiation when Eurasian ancestry is artificially blocked using a computer program. This may at first seem surprising, but it could indicate a high level of within-group diversity resulting from the large migrations and founder populations described above. Alternatively, it may represent back-migrations of relatively small bands of Eurasians who had been exposed to strong selective pressures and thus the reshaping of parts of their genomes.

The African Genome Variation Project states that part of its vision is to facilitate medical genomic studies in Africa, through generating data on African populations and also through developing capacity for genomic research on the continent. Towards this goal, the authors present a comparison of existing genome chips — devices used for identifying genetic variations that are associated with diseases — and suggest design modifications for chips that will better capture genomic variation in African populations. Such improvements in chip design will be of immediate benefit, especially to the large-scale genome projects covering different African populations that are already under way as part of the H3Africa Consortium<sup>4</sup>.

Developing countries across the globe are at disparate stages of engaging with human-genomics research. In some parts of Asia and Latin America, there has been a ready

engagement with genomics and other high-throughput technologies, as a result of the active involvement of academic institutions and governmental support. Such involvement is crucial if human genetics and genomics is to be integrated into education and training, and into other disciplines such as health and agriculture, and thus for this science to contribute to enhancing a state's development. A good example of this integration is the coordinated effort of the 100,000 Genomes Project

in the United Kingdom with Health Education England, under the auspices of the UK National Health Service. This initiative integrates workforce training designed to raise awareness of the predicted impact of the 'omics revolution' on health care.

Understandably, there has been substantial 'data-generation envy' from countries and researchers on the African continent. Although adoption of the international genomics momentum is happening at different speeds across the continent, large-scale genomic health projects such as the H3Africa Consortium, which is driven by the African Society of Human Genetics with funding from the US National Institutes of Health and Britain's Wellcome Trust, are serving as examples for networked research projects in Africa. However, effort is still needed to draw those researchers and countries not yet part of H3Africa into this 'genomic fest', and to ensure that government engagement, as well as education and training reforms, continue to grow beyond universities. ■

**Raj Ramesar** is in the MRC Human Genetics Research Unit, Division of Human Genetics, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Observatory 7925, South Africa.  
e-mail: raj.ramesar@uct.ac.za

1. Gurdasani, D. *et al.* *Nature* **517**, 327–332 (2015).
2. Lachance, J. *et al.* *Cell* **150**, 457–469 (2012).
3. Pickrell, J. K. *et al.* *Proc. Natl Acad. Sci. USA* **111**, 2632–2637 (2014).
4. The H3Africa Consortium. *Science* **344**, 1346–1348 (2014).

This article was published online on 3 December 2014.

#### PHYSICAL CHEMISTRY

## Hydrophobic interactions in context

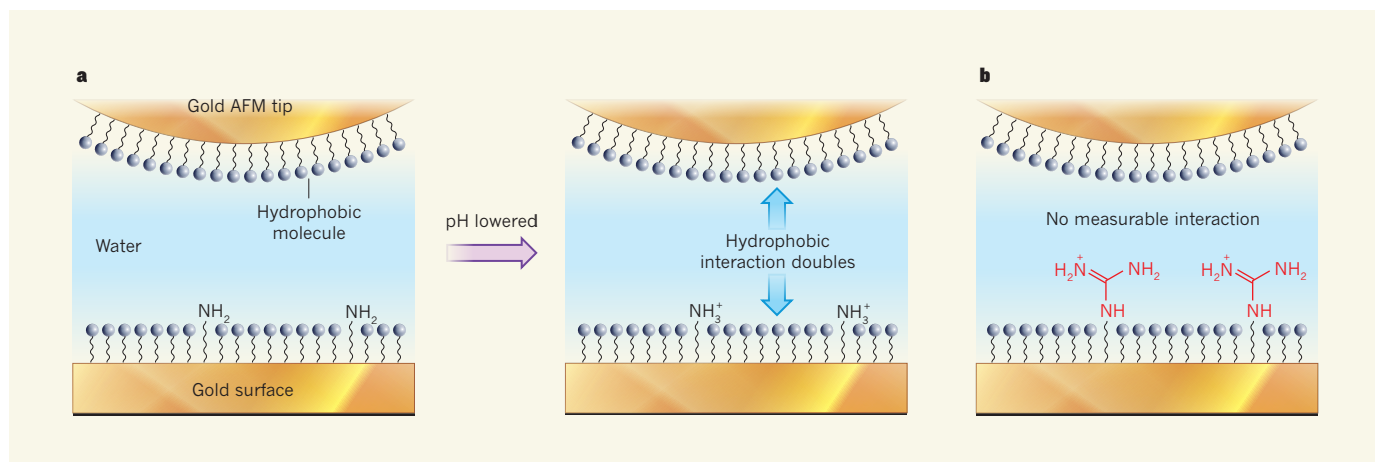
**The finding that immobilized ions can alter the strength of hydrophobic interactions between molecules suggests a strategy for tuning hydrophobicity to optimize molecular recognition and self-assembly processes. SEE LETTER P.347**

**SHEKHAR GARDE**

**O**il and water do not mix. At the molecular level, this 'de-mixing' tendency, known as the hydrophobic interaction, is thought to drive many self-assembly processes, such as protein folding, formation of micelles and membranes, and molecular recognition<sup>1</sup>. On page 347 of this issue, Ma *et al.*<sup>2</sup> present experimental evidence that the strength of hydrophobic interactions is

dramatically affected by two biologically relevant cations — ammonium and guanidinium — when these ions are immobilized near hydrophobic patches of molecules.

The oily core of a globular protein is formed by segregation of its hydrophobic groups. By contrast, protein surfaces are mosaics containing not only polar and charged groups, but also hydrophobic ones, which frequently occur in patches that are crucial for binding and recognizing other molecules, and for assembling



**Figure 1 | Context-dependent hydrophobic interactions.** Ma *et al.*<sup>2</sup> measured the hydrophobic interaction between the tip of an atomic force microscope (AFM) coated with a hydrophobic monolayer of molecules and various self-assembled monolayers of molecules on a gold surface in water. **a**, When amino ( $-\text{NH}_2$ ) groups were immobilized within nanometres of hydrophobic groups

in a monolayer, and were then charged to form ammonium groups ( $-\text{NH}_3^+$ ) by lowering the pH of the surrounding solution, the hydrophobic interaction doubled in strength. **b**, By contrast, when guanidinium groups (red, or their uncharged equivalents, guanidine groups, not shown) were incorporated into the monolayer, the hydrophobic interaction was essentially eliminated.

proteins into larger structures. These patches do not exist in isolation, but reside in the chemical and topographical context of their surroundings. There is a growing realization from simulation studies<sup>3,4</sup> that the surrounding chemical environment affects hydrophobic interactions in non-intuitive ways, but quantifying this effect experimentally has remained a challenge.

Hydrophobic interactions are affected by various factors, including temperature, pressure and the presence of various additives in solution. In particular, dissolved salt ions are free to move about and distribute themselves, depleting from or binding to interfaces between dissolved molecules and the surrounding water, and thus affect the solubility and interactions of hydrophobic species in water<sup>5</sup>. But what happens when an ion is chemically immobilized near a hydrophobic patch?

Ma and colleagues consider this question for ammonium and guanidinium ions, which are biologically relevant because they form the positively charged moieties of the amino acids lysine and arginine, respectively. The authors prepared self-assembled monolayers that present a mixture of hydrophobic groups (60%) and ionic species (either ammonium or guanidinium ions, 40%), which allowed them to position the ions within nanometres of hydrophobic regions. They then used atomic force microscopy (AFM) to measure the adhesive force between the resulting surfaces and a hydrophobic AFM tip in aqueous solutions (Fig. 1).

To tease out the component of the force attributable to hydrophobic interactions, Ma and co-workers added methanol to the solution, which eliminates most of the hydrophobic interactions. The sensitivity of their experimental systems to methanol therefore acted as a signature of the hydrophobic component of the force. They also changed the pH

of the solution to toggle ions between their charged and neutral states. This was expected to affect electrostatic interactions, but the authors found that it also altered hydrophobic interactions. The investigators compared the behaviour of the mixed monolayers with those of reference systems, such as purely hydrophobic or ionic monolayers, to draw their conclusions.

The key result is that the strength of hydrophobic interactions between the mixed monolayer and the tip roughly doubles when neutral amine groups immobilized within nanometres of hydrophobic regions are charged to form ammonium ions by lowering the pH (Fig. 1a). By contrast, similarly placed guanidinium ions essentially eliminate the hydrophobic interactions at all pH values studied (Fig. 1b). Excitingly, these observations hold not only for interactions involving the extended hydrophobic surfaces of self-assembled monolayers, but also for interactions between the hydrophobic AFM tip and hydrophobic patches on single-molecule helices of  $\beta$ -peptides that contain ammonium (lysine) or guanidinium (arginine) groups about 1 nm away from a patch.

A major strength of the work is that it presents an experimental framework for characterizing hydrophobic interactions that is versatile enough to probe how they are modulated by the proximity of other moieties. Although the authors focus on two cations here, this framework could be used to systematically study how a wide range of other groups and/or chemical patterns might modulate inherent hydrophobic interactions.

Because hydrophobic interactions are mediated by water, a deeper discussion of their context dependence will require an understanding of the role of water molecules. Previously reported work<sup>3,6</sup> has shown that density fluctuations of water near a surface serve as one robust and context-dependent molecular

measure of the hydrophobicity of the surface. Understanding how ammonium and guanidinium ions influence water's behaviour (such as its molecular packing, orientations and density fluctuations close to surfaces) may shed light on the molecular origins of the current results.

Ammonium and guanidinium ions interact with water molecules differently<sup>7</sup>. Ammonium is smaller than guanidinium and is well hydrated, and its salts are often used to stabilize or help to crystallize proteins. The guanidinium ion, with its central carbon atom attached to three amino ( $-\text{NH}_2$ ) groups, is a planar, disc-like molecule. It can donate hydrogen bonds to water molecules along its edge, but is easily depleted of water molecules along its flat faces<sup>8</sup>. This peculiar behaviour enables guanidinium ions to pair with other molecules, including with other guanidinium ions<sup>9</sup>, or with hydrophobic surfaces along its flat face, and also to denature proteins (that is, to unravel their three-dimensional structures). What part this hydration behaviour of the ions plays in the effects observed by Ma and co-workers remains to be seen.

The fact that ions in close proximity can dramatically enhance or diminish hydrophobic interactions may provide an invaluable handle for manipulating and designing protein–protein interactions. In this respect, the effects of negative ions — such as carboxylates, phosphates or sulfates — could be just as interesting and relevant as those of the positively charged ions studied by Ma *et al.*, because they are ubiquitous in naturally occurring compounds such as proteins, DNA and heparin (an anticoagulant). Indeed, when negatively charged ionic groups are inserted at the edge of a key hydrophobic binding loop of certain antibodies, the resulting molecules no longer stick to each other, but still bind to their targets<sup>10</sup>: small aggregates or fibrils of



$\beta$ -amyloid, the peptide that forms the main component of plaques found in the brains of patients with Alzheimer's disease.

More broadly, the fundamental questions about the molecular origins of the context dependence of hydrophobicity raised by Ma and colleagues' work are ripe for investigation using theory simulations and experiments. Addressing these questions is important, because the findings reassert that the different factors involved when two proteins (or two chemically heterogeneous surfaces) interact

with each other are not additive — throwing a spanner in the works of simplistic models that assume this. ■

**Shekhar Garde** is in the Department of Chemical and Biological Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180, USA.  
e-mail: [garde@rpi.edu](mailto:garde@rpi.edu)

1. Chandler, D. *Nature* **437**, 640–647 (2005).
2. Ma, C. D., Wang, C., Acevedo-Vélez, C., Gellman, S. H. & Abbott, N. L. *Nature* **517**, 347–350 (2015).

3. Patel, A. J. & Garde S. J. *Phys. Chem. B* **118**, 1564–1573 (2014).
4. Giovambattista, N., Debendetti, P. G. & Rossky, P. J. *J. Phys. Chem. C* **111**, 1323–1332 (2007).
5. Jungwirth, P. & Cremer, P. S. *Nature Chem.* **6**, 261–263 (2014).
6. Patel, A. J. *et al. J. Phys. Chem. B* **116**, 2498–2503 (2012).
7. Werner, J. *et al. J. Phys. Chem. B* **118**, 7119–7127 (2014).
8. Mason, P. E. *et al. J. Am. Chem. Soc.* **126**, 11462 (2004).
9. Shih, O. *et al. J. Chem. Phys.* **139**, 035104 (2013).
10. Perchiacca, J. M., Ladiwala, A. R. A., Bhattacharya, M. & Tessier, P. M. *Protein Eng. Design Select.* **25**, 591–602 (2012).

## NEUROSCIENCE

# Dragonflies predict and plan their hunts

**An analysis reveals that the dragonfly's impressive ability to catch its prey arises from internal calculations about its own movements and those of its target — the first example of such predictions in invertebrates. [SEE ARTICLE P.333](#)**

STACEY A. COMBES

Imagine a ballet dancer moving across the stage to meet his partner, who is leaping and pirouetting towards him. To catch her at the right moment, he must predict where she will end up and determine how he should move to intercept her. To do this, his mind anticipates how her image should grow as they move towards each other, allowing him to rapidly identify and react to unexpected changes, such as a stumble that lowers her speed. Until now, this type of complex control, which incorporates both prediction and reaction, had been demonstrated only in vertebrates. However, in this issue, Mischiati *et al.*<sup>1</sup> (page 333) show that dragonflies on the hunt perform internal calculations every bit as complex as those of a ballet dancer.

Dragonflies are formidable predators. With huge eyes that provide an almost spherical view of the world, they perch on vegetation, waiting for prey to drift overhead. When the time is right, they shoot off in pursuit, scooping up victims with their hairy legs in less than half a second (Fig. 1). Dragonflies succeed in catching their prey about 95% of the time<sup>2,3</sup>, and this prowess has been attributed to their visual acuity and lightning-quick reflexes — in particular to the specialized visual neurons that detect the motion of a target and instruct the wings to react<sup>4</sup>.

If dragonflies' pursuits were guided purely by their reactions to the movements of their prey, one would predict a one-to-one mapping between prey manoeuvres and dragonfly reactions. Mischiati and colleagues show that this is clearly not the case. Dragonflies do respond

to some prey manoeuvres, but more often they do not. And what's more, the authors report that the majority of dragonfly manoeuvres are not associated with any change in prey motion.

Some of these prey-independent manoeuvres are related to the mechanical requirements of prey capture: dragonflies align themselves with the flight path of their prey, approaching from below, most probably to reduce the likelihood of detection. Their bodies and heads move independently during prey capture, with the

head remaining locked onto its target<sup>2</sup> while the body manoeuvres into the optimal orientation for capture. Until now, it had been assumed that these target-locking head motions were performed reactively, with dragonflies moving their heads to re-centre the prey after any motion — either their own or that of their prey — that shifts the target from their sights.

To tease apart the causes and consequences of head movements during prey capture, Mischiati *et al.* performed extremely accurate, high-speed measurements of prey position, and of dragonfly head and body orientation. Such measurements are possible only in a controlled, indoor setting, where dragonflies typically refuse to chase prey. To get around this problem, the authors constructed an indoor flight arena, complete with backdrops of natural scenery and lighting that simulated a bright, sunlit day. Once they had quantified the movements of dragonfly and prey, the researchers calculated how the image of the prey moved across the dragonfly's eyes, as the result of the movements of both parties. These calculations revealed that the dragonfly's



**Figure 1 | A dragonfly in flight.** Mischiati *et al.*<sup>1</sup> found that dragonflies on the hunt make internal calculations about the movements of their prey and themselves.

head motions are remarkably effective at cancelling out the large image drift across the eye that would have resulted from its own body rotations and the prey's anticipated motion. Such cancellation ensures that the prey image remains within a few degrees of the dragonfly's visual acute zone, in which its sight is at its sharpest.

Most notably, these data show that, rather than adjusting head position after the prey image drifts outside the visual acute zone, dragonflies adjust their head positions in near-perfect synchrony with the motions that would cause image drift. This precise timing led Mischiati and co-workers to surmise that dragonflies must be generating predictions using internal models of how prey- and self-motion will affect the location of the prey image on their eyes, and moving their heads to compensate before image drift occurs.

This type of predictive control confers an advantage when compared to a purely reactive strategy. First, although the dragonfly's response time is quite fast (approximately 50 milliseconds), this still accounts for 10% to 25% of a typical chase, so reacting only after each change in prey- or self-motion would extend the duration of a chase considerably. Second, because the dragonfly's own body rotations cause much more image drift than the motion of a distant prey item, nullifying this large image drift before it occurs means that the dragonfly's visual system is more sensitive to unexpected prey manoeuvres, which it can then respond to reactively.

Of course, there is a limit to how much laboratory studies can tell us about dragonfly predation in the wild. The current experiments used either slow, laboratory-reared fruit flies that rarely take evasive action<sup>3</sup>, or artificial prey undergoing a single change in speed. So, although Mischiati and colleagues' results indicate that most manoeuvres relate to the dragonfly's pre-choreographed capture strategies, in the wild, dragonflies must contend with prey that behave more unexpectedly. Many wild insects fly erratically at all times, or detect approaching predators and perform evasive manoeuvres. In these cases, reactive control is likely to dominate the dragonfly's actions. Nonetheless, predictive steering strategies presumably still underlie such more challenging pursuits.

More broadly, Mischiati and colleagues' results open up new avenues for exploring the mechanistic basis of complex behaviours involving both predictive and reactive control. In situations such as those presented in this study<sup>1</sup>, the brain can align its internal predictions with an appropriate reaction when reality deviates from expectations. These types of behaviour — particularly the use of 'forward models', in which an animal predicts how its own actions will affect its sensory feedback — had previously been demonstrated only in vertebrates<sup>5–7</sup>, in which analysis of neural

circuitry is challenging. By contrast, dragonflies have accessible neural circuitry, and their relatively large size allows for measurements of behaviour and neural activity during free flight. Hunting dragonflies thus present a rare opportunity for conducting detailed, mechanistic studies of the neural circuits that underlie complex behaviours. ■

**Stacey A. Combes** is in the Department of Organismic and Evolutionary Biology, Concord Field Station, Harvard University, Bedford, Massachusetts 01730, USA.  
e-mail: scombes@oeb.harvard.edu

1. Mischiati, M. *et al.* *Nature* **517**, 333–338 (2015).
2. Olberg, R. M., Seaman, R. C., Coats, M. I. & Henry, A. F. J. *Comp. Physiol. A* **193**, 685–693 (2007).
3. Combes, S. A., Rundle, D. E., Iwasaki, J. M. & Crall, J. D. *J. Exp. Biol.* **215**, 903–913 (2012).
4. Gonzalez-Bellido, P. T., Peng, H., Yang, J., Georgopoulos, A. P. & Olberg, R. M. *Proc. Natl Acad. Sci. USA* **110**, 696–701 (2012).
5. Wolpert, D. M., Ghahramani, Z. & Jordan, M. I. *Science* **269**, 1880–1882 (1995).
6. Mehta, B. & Schaal, S. *J. Neurophysiol.* **88**, 942–953 (2002).
7. Webb, B. *Trends Neurosci.* **27**, 278–282 (2004).

This article was published online on 10 December 2014.

## ORGANIC CHEMISTRY

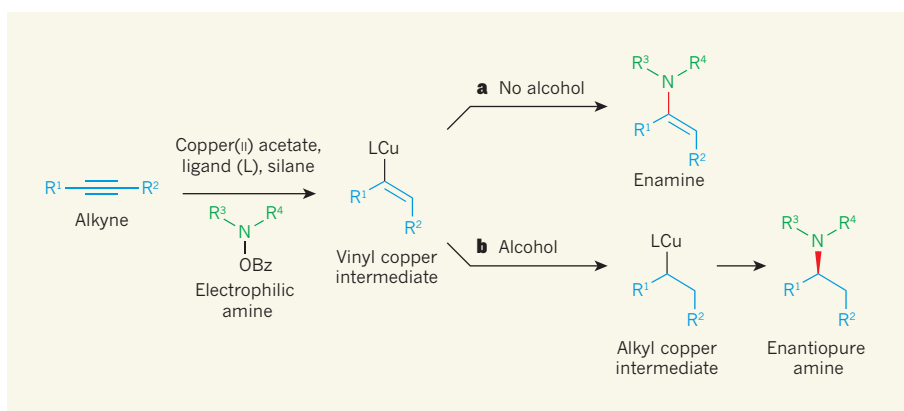
# One catalyst, two reactions

**A catalyst has been tuned to make different compounds from the same molecules in carbon–nitrogen bond-forming reactions, depending on the conditions used. The products are potential building blocks for biologically active molecules.**

EMMANUELLE SCHULZ

The ability to readily synthesize structurally complex molecules containing nitrogen atoms is crucial for organic chemists because such compounds have widespread applications, for example as drugs. But the nitrogen atoms must be incorporated into molecules at particular locations with respect to other atoms, using methods that are compatible with chemical groups already present in those molecules. The three-dimensional

arrangement of atoms must also be mastered to prepare 'enantiopure' compounds (single mirror-image isomers of compounds, called enantiomers), rather than a one-to-one mixture of enantiomers that must then be tediously separated. This is crucial for medicinal chemists, because different enantiomers can have different, sometimes even opposing, biological activities. Writing in *Nature Chemistry*, Shi and Buchwald<sup>1</sup> report a variant of a carbon–nitrogen bond-forming reaction that solves many of the problems associated



**Figure 1 | Selective hydroamination reactions of alkynes.** Shi and Buchwald<sup>1</sup> report that a vinyl copper intermediate forms when an alkyne reacts with an electrophilic amine in the presence of copper(II) acetate, a ligand molecule (which binds to the copper ions) and a silane ( $\text{HSiCH}_3(\text{OC}_2\text{H}_5)_2$ ). **a**, In the absence of an alcohol, the intermediate reacts with the electrophilic amine to form an enamine. The carbon–nitrogen bond formed during the reaction is shown in red. **b**, But in the presence of an alcohol, an alkyl copper intermediate forms through a cascade of reactions, and produces amine products in enantiopure form (as single mirror-image isomers).  $\text{R}^1$  to  $\text{R}^4$  represent general chemical groups; Bz is a benzoyl group.



with making organic nitrogen-containing compounds.

There are many ways of synthesizing carbon–nitrogen (C–N) bonds. Some are based on substitution reactions, in which an atom or group of atoms is displaced to form the desired C–N connection. Such reactions typically provide high product yields but generate huge quantities of chemical waste, which is unsustainable given the need for environmentally friendly, ‘green’ chemistry<sup>2</sup>. Addition reactions, in which all the atoms present in an incoming molecule are incorporated into the target compound, are therefore preferred.

The hydroamination reaction<sup>3</sup> perfectly illustrates this second type of transformation. Its most direct form involves the addition of a nitrogen compound to carbon–carbon (C–C) double or triple bonds, generating both a C–N and a carbon–hydrogen (C–H) connection without any by-products (Fig. 1). But the reaction is not straightforward, mainly because electrons in the two reactants repel each other. It is therefore crucial to use a catalyst to couple the reactants together.

Because hydroamination has been so widely studied, numerous catalysts based on virtually all the elements of the periodic table have been reported for this reaction. It is possible to obtain products with high enantiopurity, but there is not yet a universal method that is applicable to all molecules<sup>4</sup>. The reactions frequently require catalysts based on precious and expensive metals, harsh conditions (such as high temperatures) and chemically activated reactants, or they may be sensitive to factors such as air or moisture. They therefore do not meet the requirements of modern synthesis.

Shi and Buchwald report a straightforward solution to these difficulties. They used a simple copper salt as a catalyst to perform hydroamination reactions between readily available alkynes (compounds with a C–C triple bond) and nitrogen-containing compounds known as electrophilic amines that have a specific affinity for the electronic character of alkyne triple bonds. Starting from the same reactants, the authors found that by changing the reaction conditions, they could easily and preferentially prepare either one of two compounds — an enamine or an amine — possessing a highly desirable C–N bond.

By combining an alkyne, an electrophilic amine, a copper salt, a ligand (an enantiopure phosphorus-containing molecule that binds to the reactive copper centre of the catalyst) and a silane compound (which provides hydrogen atoms needed for the C–H bond formation that occurs during the reaction), Shi and Buchwald observed the exclusive formation of enamines through an intermediate known as a vinyl copper species (Fig. 1). But when the authors simply added an alcohol (a source of hydrogen ions, H<sup>+</sup>) to the mixture, no enamine formed; instead, the vinyl copper species underwent a cascade of reactions, ultimately

and exclusively forming an alkyl copper intermediate. They then used this to prepare amines through the formation of new C–N bonds, as had previously been reported simultaneously by Buchwald’s group<sup>5</sup> and by others<sup>6</sup>. Crucially, the isolated products are almost completely enantiopure.

Shi and Buchwald have achieved a remarkable feat: by taming the reactivity of all the intermediate species, they performed reactions chemoselectively (forming either an enamine or an amine), regioselectively (grafting the nitrogen atom only to a targeted carbon atom) and, in the case of the amine formation, enantioselectively. The authors went on to show that the reactions are applicable to a broad range of substrates, and that they work under ‘mild’ reaction conditions that tolerate the presence of a wide range of chemical groups. Finally, the authors demonstrated the utility of their reactions by using them to transform a variety of easily prepared compounds into marketed pharmaceutical compounds — in one case, in nearly enantiopure form.

This work represents a big step forward for the highly desirable hydroamination reaction.

## HIV

## Seeking ultimate victory

**HIV variants that have mutated to escape T-cell immune responses dominate the latent viral reservoir in most patients on antiretroviral therapy. This finding will need to guide therapeutic approaches targeting reactivated virus. [SEE LETTER P.381](#)**

LOUIS J. PICKER & JEFFREY D. LIFSON

In the ancient treatise *The Art of War*<sup>1</sup>, Sun Tzu advises would-be victorious generals to “know thy enemy”. Perhaps nowhere in medicine is this admonition more apt than in the effort to cure HIV infection. Although both physicians and patients would enthusiastically welcome a simple, virus-directed therapy that could safely and efficiently purge HIV from the body, the biological obstacles to HIV cure — either viral eradication or complete, sustained, off-treatment remission — are too many and too complex to permit such a straightforward approach<sup>2,3</sup>. A paper by Deng *et al.*<sup>4</sup> on page 381 of this issue contributes to the effort to defeat HIV by rigorously documenting one of these obstacles to cure, while at the same time pointing to a possible strategy to overcome this particular enemy strength.

Although combination antiretroviral therapy (cART) is highly effective at suppressing active HIV replication, these drugs do not target the reservoirs of latent (non-replicating) HIV

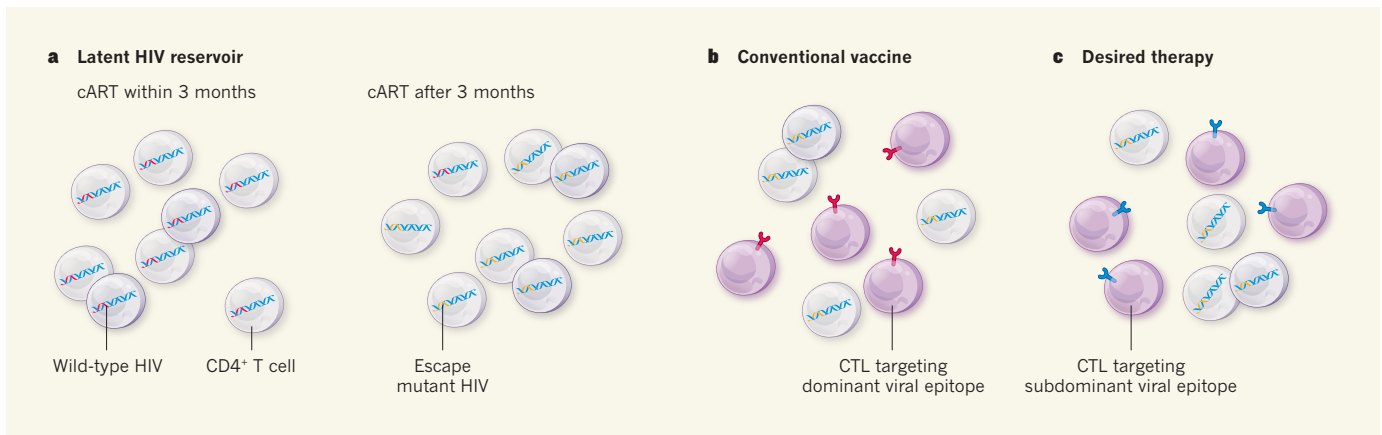
Its strength lies in the variety of products that can be obtained by a simple change in the reaction conditions. Nevertheless, there is still room for improvement. In particular, the chemical groups that decorate the starting alkyne can affect the progress and outcome of the reaction. The search for universal catalysts is not yet over, but Shi and Buchwald’s discovery, based on the use of cheap and non-toxic copper salts, will doubtless inspire further progress. ■

**Emmanuelle Schulz** is at the *Institut de Chimie Moléculaire et des Matériaux d’Orsay, Université Paris-Sud, 91405 Orsay Cedex, France.*  
e-mail: [emmanuelle.schulz@u-psud.fr](mailto:emmanuelle.schulz@u-psud.fr)

1. Shi, S.-L. & Buchwald, S. L. *Nature Chem.* **7**, 38–44 (2015).
2. Anastas, P. & Eghbali, N. *Chem. Soc. Rev.* **39**, 301–312 (2010).
3. Müller, T. E., Hultsch, K. C., Yus, M., Foubelo, F. & Tada, M. *Chem. Rev.* **108**, 3795–3892 (2008).
4. Hannedouche, J. & Schulz, E. *Chem. Eur. J.* **19**, 4972–4985 (2013).
5. Zhu, S., Niljianskul, N. & Buchwald, S. L. *J. Am. Chem. Soc.* **135**, 15746–15749 (2013).
6. Miki, Y., Hirano, K., Satoh, T. & Miura, M. *Angew. Chem. Int. Edn* **52**, 10830–10834 (2013).

that are established in infected CD4<sup>+</sup> memory T cells (a type of immune cell) early in infection and maintained over the lifetime of an individual<sup>2,3</sup>. These latently infected cells, which do not express viral proteins and are thus invisible to surveillance by the host’s immune system, remain a source of virus that can reignite progressive infection when cART is stopped.

Drugs have been identified that can induce viral expression from latently infected cells *in vitro*, and possibly *in vivo*, and although overcoming latency remains a challenge, it looks increasingly likely that such agents will be improved to the point at which viral reactivation can be substantially accelerated<sup>5,6</sup>. However, contrary to initial hopes, cells with reactivated virus do not invariably die as a result of cell-lytic viral replication, but instead may survive and possibly even proliferate<sup>7</sup>, indicating that eliminating such cells is an important therapeutic goal. The obvious solution to this problem is to co-opt the host immune system, particularly the ability of CD8<sup>+</sup> T cells (also called cytotoxic T lymphocytes, or CTLs) to



**Figure 1 | Immune escape complicates targeting of the HIV reservoir.**

**a**, A reservoir of CD4<sup>+</sup> T cells containing latent (non-replicating) viruses is established soon after infection with HIV. Deng *et al.*<sup>4</sup> show that, in patients who have commenced combined antiretroviral therapy (cART) within three months of infection, most of these latent viruses have wild-type sequences at the regions encoding the viral structures towards which the CD8<sup>+</sup> cytotoxic T cells (CTLs) of the host immune system are predominantly directed (dominant viral epitopes; red). However, in patients starting cART later than this, more than 98% of the viruses have already mutated these sequences to

‘escape’ the immune response (dominant-epitope escape mutants; orange).

Viral sequences encoding structures towards which little CTL activity is directed (subdominant viral epitopes) are shown in blue. **b**, The findings suggest that, in patients who start cART after three months, conventional vaccines will be ineffective in combating viruses reactivated from this reservoir, because such vaccines primarily mobilize CTLs that target the wild-type dominant viral epitopes. **c**, Instead, vaccines will need to be designed that mobilize CTLs targeting subdominant (non-mutated) viral epitopes and that may thus be able to kill cells harbouring reactivated latent virus.

recognize and kill infected cells that exhibit any HIV-gene expression. But even here, perhaps not surprisingly, the virus seems to retain the upper hand, by using its capacity for extensive genetic variation and rapid evolution to evade CTLs, through a process known as mutational escape<sup>8</sup>.

Deng and colleagues addressed the key question of whether reactivated latent virus is recognized by an infected individual’s HIV-specific T cells. The answer, at least in individuals started on cART more than three months after infection (chronic-phase cART), was daunting: more than 98% of HIV integrated into the patients’ latently infected CD4<sup>+</sup> T cells contained sequence changes in regions (epitopes) of the viral Gag protein (the primary target of effective CTL responses) corresponding to previously characterized CTL-escape mutants (Fig. 1a). The researchers confirmed that these Gag mutations conferred functional immune escape by demonstrating that CD8<sup>+</sup> T cells from some of the patients given chronic-phase cART recognized the wild type, but not the mutant, forms of the relevant Gag epitopes. They also verified that the mutant sequences were present in replication-competent virus derived from the chronic-phase cART patients’ latently infected CD4<sup>+</sup> T cells, indicating that mutant virus could contribute to viral re-emergence if cART were stopped.

By contrast, the authors observed that, in patients who began cART within the first three months after HIV infection, the latent virus was of predominantly wild-type sequence (Fig. 1a). Intriguingly, these findings suggest that the latent HIV reservoir in untreated subjects is more dynamic than was previously thought. Of greater clinical importance is the implication that the conventional CTL

responses against dominant HIV epitopes (which arise during the early phase of HIV infection and help to determine the level of viral replication during the chronic phase) will almost certainly not contribute to destroying reactivated latently infected cells (Fig. 1b). As such, these conventional CTL responses cannot be expected to supply the cell killing required for effective cure, even if the response is boosted by therapeutic vaccination using conventional vaccines — at least for the vast majority of HIV-infected individuals who are started on cART during chronic infection.

But the news from Deng and colleagues is not all bad. The investigators also found that chronically infected patients on cART retain CD8<sup>+</sup> T cells that are specific for subdominant, non-mutated Gag epitopes, and show that these cells can recognize and even kill cells infected with mutated virus *in vitro* and *in vivo* (in a humanized mouse model). The ‘rub’ with these data is that the efficiency of this cell killing seems to depend on *in vitro* preactivation of the CD8<sup>+</sup> T cells, and Deng *et al.* provide no evidence that these broadly targeted CD8<sup>+</sup> T cells can be directly recruited *in vivo* to clear cells harbouring reactivating virus in cART-suppressed infections.

However, the data are important for their clear indication that exploiting CTL responses in attempts to cure infection will require approaches that induce CD8<sup>+</sup> T-cell responses to subdominant epitopes (Fig. 1c) or to epitopes that are not naturally targeted at all during the course of HIV infection, as has been described for SIV infections in rhesus macaques<sup>9</sup>. Even these approaches will need to surmount yet more hurdles, including the presence of residual virus in immune-privileged ‘sanctuary’

sites, such as B-cell follicles<sup>10</sup>.

Sun Tzu also advises generals to “avoid what is strong, and strike at what is weak”. For curing HIV infections, this advice would translate to the identification and therapeutic targeting of HIV’s weaknesses, and suggests that understanding barriers to cure and defining the biology underlying these barriers are the first steps towards overcoming this viral enemy. HIV cure will almost certainly require a multimodal therapeutic approach incorporating both pharmacological activation of latent reservoirs and immune-mediated clearance mechanisms, with each component designed to exploit one of this formidable enemy’s few weaknesses. ■

**Louis J. Picker** is at the Vaccine and Gene Therapy Institute, Oregon Health & Science University, Beaverton, Oregon 97006, USA.

**Jeffrey D. Lifson** is in the AIDS and Cancer Virus Program, Leidos Biomedical Research, Inc., Frederick National Laboratory, Frederick, Maryland 21702, USA.

e-mails: pickerl@ohsu.edu;

lifsonj@mail.nih.gov

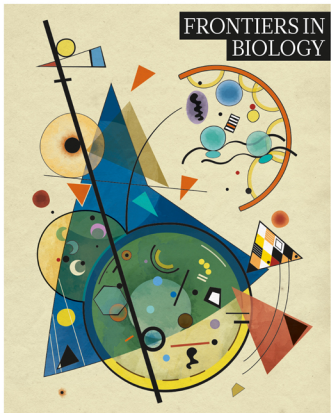
1. Tzu, S. *The Art of War* (Ulysses, 2007).
2. Katlama, C. *et al.* *Lancet* **381**, 2109–2117 (2013).
3. Barouch, D. H. & Deeks, S. G. *Science* **345**, 169–174 (2014).
4. Deng, K. *et al.* *Nature* **517**, 381–385 (2015).
5. Margolis, D. M. & Hazuda, D. J. *Curr. Opin. HIV AIDS* **8**, 230–235 (2013).
6. Thornhill, J., Fidler, S. & Frater, J. *Curr. Opin. Infect. Dis.* <http://dx.doi.org/10.1097/QCO.000000000000123> (2014).
7. Shan, L. *et al.* *Immunity* **36**, 491–501 (2012).
8. Picker, L. J., Hansen, S. G. & Lifson, J. D. *Annu. Rev. Med.* **63**, 95–111 (2012).
9. Hansen, S. G. *et al.* *Science* **340**, 1237874 (2013).
10. Fukazawa, Y. *et al.* *Nature Med.* (in the press).

This article was published online on 7 January 2015.



# natureINSIGHT

FRONTIERS IN  
BIOLOGY





**Cover illustration**  
Nik Spencer

**Editor, *Nature***  
Philip Campbell

**Publishing**  
Richard Hughes

**Production Editor**  
Jenny Rooke

**Art Editor**  
Nik Spencer

**Sponsorship**  
Reya Silao

**Production**  
Ian Pope

**Marketing**  
Steven Hurst

**Editorial Assistant**  
Melissa Rose

The Macmillan Building  
4 Crinan Street  
London N1 9XW, UK  
Tel: +44 (0) 20 7833 4000  
e: [nature@nature.com](mailto:nature@nature.com)



nature publishing group

**T**he *Nature* Insight 'Frontiers in Biology' aims to cover timely and important developments across biology, ranging from subcellular molecular mechanisms to whole-organism physiology, and biomedicine.

The Insight begins with a review on the amygdala, a brain structure that responds to both negative and positive associations during functionally different behaviours. This structure has assumed a complex position within the circuits underlying emotional and motivational processes. Embracing this complexity, Patricia Janak and Kay Tye reveal how advances in genetic targeting, anatomical tracing and neuronal activity modulation are providing the means to define a road map for this important structure's complex connectivity.

David Artis and Hergen Spits then go on to review the rapidly moving field of innate lymphocyte biology. Innate lymphoid cells are the most recently identified constituents of the innate immune system. They have important roles in protective immunity and inflammation, integrating innate and adaptive immune responses and controlling tissue homeostasis in infection, chronic inflammation, metabolic disease and cancer.

The ability to sense and to respond to nutrients such as glucose, amino acids and lipids is essential for life. David Sabatini and colleagues explore how mammals sense these nutrients and discuss the deregulation of sensing mechanisms in disease.

Programmed cell death is essential for many physiological processes, including the shaping of developing organs, epithelial cell renewal and lymphocyte selection. Manolis Pasparakis and Peter Vandenabeele discuss the regulation, initiation and execution of different types of regulated cell death with a particular focus on the non-apoptotic cell-death process necroptosis.

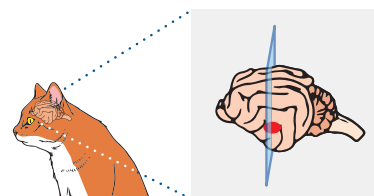
DNA methylation is an epigenetic modification that is generally associated with gene silencing. Recent technological advances have enabled the generation of genomic maps at unprecedented resolution that should help resolve its specific functions. In the final Review, Dirk Schübeler summarizes our current understanding of the regulation and function of DNA methylation, and discusses its utility as a cellular marker in basic biology and biomedicine.

**Alex Eccleston, Noah Gray, Sadaf Shadan & Ursula Weiss**  
*Senior Editors*

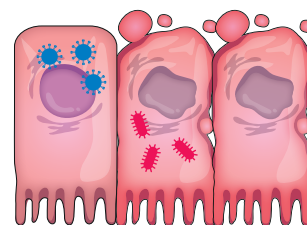
### CONTENTS

#### REVIEWS

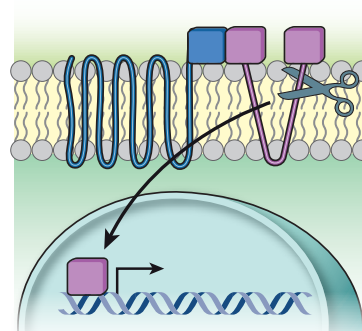
- 284 From circuits to behaviour in the amygdala**  
*Patricia H. Janak & Kay M. Tye*



- 293 The biology of innate lymphoid cells**  
*David Artis & Hergen Spits*



- 302 Nutrient-sensing mechanisms and pathways**  
*Alejo Efeyan, William C. Comb & David M. Sabatini*



- 311 Necroptosis and its role in inflammation**  
*Manolis Pasparakis & Peter Vandenabeele*

- 321 Function and information content of DNA methylation**  
*Dirk Schübeler*



# From circuits to behaviour in the amygdala

Patricia H. Janak<sup>1,2</sup> & Kay M. Tye<sup>3</sup>

**The amygdala has long been associated with emotion and motivation, playing an essential part in processing both fearful and rewarding environmental stimuli. How can a single structure be crucial for such different functions? With recent technological advances that allow for causal investigations of specific neural circuit elements, we can now begin to map the complex anatomical connections of the amygdala onto behavioural function. Understanding how the amygdala contributes to a wide array of behaviours requires the study of distinct amygdala circuits.**

**A**lthough humans possess a number of cognitive abilities that differentiate us from other animals, we share emotional behaviours — defined as behavioural responses to emotionally significant stimuli such as food or threats — with other vertebrates. The amygdala is a brain region that is important for emotional processing, the circuitry and function of which has been well-conserved across evolution (Fig. 1), although species differences do exist<sup>1</sup>. Even non-mammalian species such as reptiles, birds and fish have an amygdala-like brain region with similar circuits and functions to the amygdala in mammals<sup>2–5</sup>.

Conservation of amygdala circuitry allows findings from one species to inform our appreciation of amygdala functioning in others. Understanding the intricacies of amygdala circuitry is of tremendous importance given that the amygdala is implicated in a wide range of disease states, including addiction, autism and anxiety disorders. Regarding forward translation, features identified using new approaches for neural circuit dissection in rodents have the potential to be directly relevant to humans in well-conserved structures such as the amygdala. For reverse translation, observed correlations between amygdala function and human behaviour can then be brought back to animal models to facilitate the elucidation of underlying mechanisms using systematic, iterative experimentation. The field is ripe for this type of translation, as reflected by the growing emphasis on amygdala research across species (Fig. 2). The collective body of work supports a view of the amygdala as a composite of parallel circuits that affect multiple aspects of emotional behaviour. This Review focuses on recent advances that have been enabled by technologies primarily used in rodents. Readers are directed to recent reviews for in-depth information on the amygdala and fear<sup>6</sup>, reward<sup>7–9</sup>, learning-related plasticity<sup>10,11</sup>, and anatomical or physiological characteristics<sup>6,12</sup>.

## First clues to the role of the amygdala in behaviour

Lesion studies in non-human primates presented the earliest clues that the amygdala was important for emotional reactions to stimuli. In 1888, Brown and Schäfer performed a bilateral ablation of the temporal lobe of a rhesus monkey (*Macaca mulatta*)<sup>13</sup>, reporting that: “A remarkable change is ... manifested in the disposition of the Monkey ... He gives evidence of hearing, seeing, and of the possession of his senses generally, but it is clear that he no longer clearly understands the meaning of the sounds, sights, and other impressions that reach him.” In addition, Brown and Schäfer, and later Klüver and Bucy<sup>14</sup>, reported reduced aggression, fear and defensive behaviours. Although these lesion effects were intermingled with features that we now attribute to other brain regions, these were the

first reports describing the role of a brain region that is important for connecting stimuli with their emotional meaning. In 1956, Weiskrantz<sup>15</sup> lesioned the amygdala, demonstrating an impairment in acquiring behavioural responses to shock-predictive cues, concluding that: “the effect of amygdectomy ... is to make it difficult for reinforcing stimuli, whether positive or negative, to become established or to be recognized as such.”

Following these early studies, amygdala lesions in both rodents<sup>16,17</sup> and humans<sup>18,19</sup> revealed a strong conservation of function across species, most notably an impairment in the recognition of fearful stimuli, and in a type of emotional learning called fear conditioning. Fear is typically studied in the laboratory using a form of associative learning known as Pavlovian conditioning, in which an initially neutral conditioned stimulus (CS) is paired with an aversive unconditioned stimulus (US), such as a footshock, leading the experimental subject to display behavioural signs of fear. This simple behavioural procedure provides a window into basic Pavlovian learning mechanisms that act to enhance survival, and is intimately tied to amygdala function.

## What is the amygdala?

Although modest in size, the amygdala is comprised of multiple interconnected nuclei nestled deep in the temporal lobe (Fig. 1). Here we focus on the basolateral complex of the amygdala (BLA; made up of the lateral (LA), basal (BA) and basomedial (BM) cell groups) and the central nucleus of the amygdala (CeA; made up of a lateral (CeL) subdivision and a medial (CeM) subdivision). The BLA consists of glutamatergic principal neurons and inhibitory interneurons. CeA neurons are primarily GABAergic, with the CeL projecting to the CeM. An interconnected sheath of GABAergic neurons, termed the intercalated cells, is also found interposed between the BLA and CeA, providing an important source of inhibition<sup>12,20</sup>.

A highly simplified view of information flow through the amygdala is as follows. The amygdala receives information about the external environment from the sensory thalamus and sensory cortices, which project strongly to the LA. The LA projects within the BLA to the BA and BM, as well as to the neighbouring CeA. The BLA is reciprocally connected with cortical regions, especially the midline and orbital prefrontal cortices (PFCs), and the hippocampus (HPC), as well as sensory association areas<sup>1</sup>; in primates, these reciprocal connections extend to primary sensory areas<sup>21</sup>. Hence, the BLA transmits information widely throughout cortical regions, but its neuronal processing is greatly affected by excitatory projections from these regions. The relative enlargement of the BLA

<sup>1</sup>Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, Maryland 21218, USA. <sup>2</sup>Department of Neuroscience, Johns Hopkins University, Baltimore, Maryland 21205, USA. <sup>3</sup>Department of Brain and Cognitive Sciences, Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

compared with the CeA between rodents and primates might result from the substantial increase in the size of cortical regions that communicate with the BLA in primates<sup>22</sup>. Predominantly unidirectional outputs of the BLA include the striatum, especially the nucleus accumbens (NAc), and the bed nucleus of the stria terminalis (BNST) and the CeA. In turn, the striatum, BNST and CeA have been considered to mediate the translation of BLA signals to behavioural output. Of note, there are exceptions to this serial model of information flow; the BA and CeA also receive sensory inputs, and the CeA contributes to some behavioural processes independently from the BLA<sup>23,24</sup>.

### From rodents to humans and back again

The neural circuits underlying Pavlovian fear conditioning have predominantly been explored in rodents. Simple tone–shock pairings produced robust and reproducible changes in amygdala neural responses to the tone<sup>25,26</sup>, such that neuronal spiking tracks the acquisition and extinction of fear behaviour in response to the tone. Human studies using functional magnetic resonance imaging (fMRI) adapted these simple fear-conditioning tasks and showed that the human amygdala is also activated by fear-conditioned stimuli, and that this activation wanes on extinction<sup>27,28</sup>. Furthermore, ventromedial prefrontal cortex (vmPFC)–amygdala circuitry was found to mediate fear extinction in both rats<sup>29,30</sup> and humans<sup>31</sup>. By applying tasks designed for rodents to humans, the circuits mediating fear acquisition and extinction were shown to be well conserved.

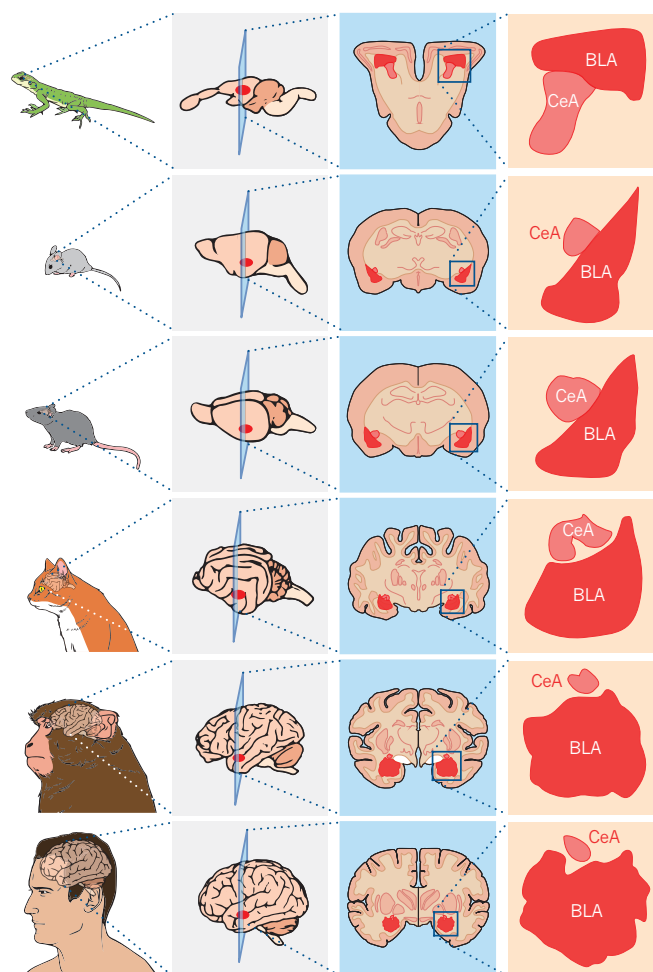
Not only does fear conditioning activate both the rodent and human amygdala, but these animals also share long-term memory processes for fear memory. When consolidated fear memories of rodents are reactivated during retrieval, they become labile, thereby requiring ‘reconsolidation’<sup>32</sup>. During this labile state, memories can be modified by interfering with the reconsolidation process using protein-synthesis inhibitors<sup>32</sup> or with the non-invasive presentation of non-fearful information in mice, rats and humans<sup>33,34</sup>.

Reverse translational approaches have become more frequent with the advent of new circuit mapping and manipulation technologies in rodents. One example in which a correlation in humans was taken to causation in rodents comes from fMRI studies in people with generalized anxiety disorder<sup>35</sup>. The observation of abnormal functional connectivity between the BLA and the CeM in these people inspired the application of optogenetic tools to probe specific projections within mouse amygdala. Optogenetics allows for the rapid and reversible activation or inhibition of neurons by directing light towards neural elements that have been artificially induced to express light-sensitive opsins. In the first demonstration of optogenetic projection-specific manipulation in a freely moving animal, the increase or decrease of transmission between the BLA and the CeA was shown to cause the reduction or augmentation of anxiety-related behaviour<sup>36</sup>, complementing the fMRI findings.

Any attempt to define the behavioural functions of amygdala neuronal activity is confronted by the dense interconnections among amygdala nuclei and between amygdala nuclei and other brain regions, and by the lack of a predictable distribution of functional cell types. The availability of new neurotechnologies and approaches is overcoming these hurdles and accelerating the mapping of function onto amygdala circuitry, revealing a complex picture of amygdalar control of behaviour. These findings, achieved using optogenetic and pharmacogenetic activation or inhibition, in conjunction with behaviour and electrophysiology, reveal causal relations between amygdala cell types and projections in various behaviours (Fig. 3). As we discuss, the behavioural functions of afferent and efferent projections had not been determined at this level of specificity until the application of these new technologies for circuit manipulation.

### New insights into circuitry for fear

The investigation of the neural basis of fear learning and expression led to the view that the amygdala is a rapid detector of aversive environmental stimuli and situations, producing affective or behavioural states to allow for adaptive responses to potential threats<sup>37,38</sup>. Along

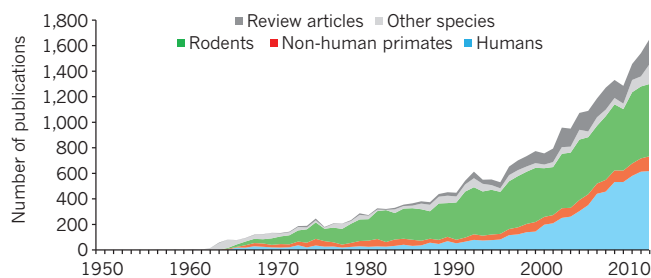


**Figure 1 | Evolution of the amygdala across species.** Primary amygdalar nuclei and basic circuit connections and function are conserved across species. An enlarged image of the basolateral complex of the amygdala (BLA) and central nucleus of the amygdala (CeA) or analogues are shown next to a coronal section from the brains of a lizard, mouse, rat, cat, monkey and human.

the way, this line of research uncovered crucial brain mechanisms of associative conditioning, arguably providing us with our best understood neurobehavioural model of learning.

The LA has been the focus of many studies because it has ready access to information about the auditory cue used in conditioning, and lesions of this region block acquisition of conditioned freezing<sup>16,17,39</sup>. LA neurons develop and maintain excitatory neural responses to the onset of an auditory cue that has been paired with a footshock US<sup>25,26,40,41</sup>. These responses *in vivo* probably arise from potentiation of sensory inputs onto LA neurons because CS–US pairings enhance measures of excitatory synaptic plasticity *in vivo*<sup>42</sup> and in acute amygdala slice preparations<sup>43–45</sup>. In this model, an initially weak afferent carrying sensory information about the CS and a strong afferent carrying US information converge onto individual principal neurons in the LA and, through a Hebbian plasticity mechanism, lead to enhanced strength of the excitatory synapses carrying CS information. This experience-dependent synaptic strengthening allows the presentation of the CS alone to activate LA neurons. The model of CS and US convergence has been explored by studies that took advantage of the temporal specificity of optogenetics to activate neural elements in a time window corresponding to the few seconds of CS or US presentation. The first study expressed the excitatory opsin, channelrhodopsin 2 (ChR2), in LA principal neurons to allow for rapid and reversible activation of these neurons during behaviour. When paired with an auditory CS, this simultaneous photoactivation of LA neurons could be used as a substitute for the footshock US, resulting





**Figure 2 | Number of studies on the amygdala.** Publications on the amygdala indexed on PubMed between 1950 and 2013 demonstrate the growing interest in amygdala research.

in conditioned freezing<sup>46</sup>. The second study showed that brief photoactivation of ChR2-expressing LA axonal terminals from the auditory thalamus (the medial geniculate nucleus, MGN) and the auditory cortex (AC) can substitute for a tone CS when paired with footshock, resulting in conditioned freezing and synaptic potentiation<sup>47</sup>.

Lesions of the CeA block expression of conditioned fear<sup>17,48,49</sup>, leading to the hypothesis that projection routes from the LA to the CeA could allow for conditioned freezing behaviour. Notably, the LA does not project directly to the CeM, the proposed output nucleus that projects to non-amygdala regions mediating behavioural and autonomic signs of fear<sup>37,50,51</sup>. The advent of cell-type- and projection-specific manipulation shed light on information flow from the LA to the CeA, and then within the CeA itself in terms of conditioned fear — information that was unattainable from lesions that remove the entire network. An important step was identifying two subpopulations of neurons within the CeL that have opposing functions and distinct genetic markers. These are neurons in the CeL that are inhibited in response to CS following fear conditioning and express protein kinase C (PKC) $\delta$ , termed CeL<sub>OFF</sub> cells, and neurons that are excited by CS following fear conditioning that are PKC $\delta^-$ , termed CeL<sub>ON</sub>. Although both subtypes of CeL neurons inhibit one another and both project to the CeM<sup>52,53</sup>, CeL<sub>ON</sub> neurons respond to a fear CS at a shorter latency than CeL<sub>OFF</sub><sup>52</sup>, suggesting that conditioned fear responses occur following activation of CeL<sub>ON</sub> neurons that inhibit CeL<sub>OFF</sub> neurons projecting to CeM output neurons<sup>53</sup>, thereby promoting freezing through disinhibition<sup>52,53</sup>.

These studies did not, however, directly address the issue of transmission between the LA and the CeA. A recent series of studies capitalized on the advantages of genetic targeting in mice to characterize a neuronal population in the CeL that receives LA input and might be involved in fear conditioning. Using a transgenic mouse line in which Cre recombinase is expressed in somatostatin-positive (SOM<sup>+</sup>) neurons, the ability of ChR2-expressing LA terminals to activate these cells could be determined. The LA was shown to form functional excitatory synapses on SOM<sup>+</sup> and PKC $\delta^-$  cells in the CeL (CeL:SOM<sup>+</sup>), and the excitatory strength of these synapses greatly increased as a result of fear conditioning<sup>54</sup>. This finding is notable because the LA was considered to be the primary site for learning-related plasticity underlying fear conditioning. The presence of experience-dependent potentiation in the CeL indicates that this region is also important for acquisition of the CS–US association, as suggested by earlier studies<sup>17,40,52</sup>, rather than merely serving as a relay of information from the LA.

The importance of this experience-dependent plasticity onto CeL:SOM<sup>+</sup> neurons in fear acquisition was demonstrated using a pharmacogenetic approach to reversibly silence only the CeL:SOM<sup>+</sup> neurons during the acquisition phase of fear conditioning. Using a cre-dependent virus to express the inhibitory DREADD receptor hM4Di, an engineered G-protein-coupled receptor activated by its exogenous ligand clozapine *N*-oxide, the SOM<sup>+</sup> cells could be reversibly and selectively inhibited. This inhibition prevented fear acquisition and excitatory synaptic enhancement<sup>54</sup>. Using the retrograde tracer cholera toxin B injected into SOM<sup>+</sup> terminal regions to allow whole cell recording from projection-defined

cells, it was revealed that SOM<sup>+</sup> cells with learning-induced synaptic enhancement project directly to the periaqueductal grey (PAG) as well as the paraventricular nucleus of the thalamus (PVT)<sup>55</sup>, bypassing the CeM. Thus, CeL projections to the CeM might not be the only CeL projections that contribute to freezing behaviour after fear conditioning. This evidence revises the view of the CeM as the primary output station of the amygdala and indicates that information at various stages of processing in the CeA is sent to the PAG.

Although pieces of information are continually emerging, these circuit analyses reveal a rich interaction between the LA and CeL — and among neurons within the CeL — in both acquiring and expressing conditioned freezing in response to a fear cue, as well as multiple neural projections from the CeL that may regulate this behavioural response. Fear conditioning to the environment in which footshock is experienced is also mediated by the amygdala. Circuit approaches to contextual fear conditioning are underway; a recent study found that optogenetic inhibition of the BLA terminals in the entorhinal cortex (EC) impairs the acquisition of contextual fear conditioning<sup>56</sup>, whereas photoexcitation of oxytocin (OT)-releasing fibres from the hypothalamus to the CeL suppresses expression<sup>57</sup>.

### Beyond fear in the amygdala

The robustness of fear conditioning provided a highly replicable behavioural paradigm that allowed researchers to delve deep into the circuit mechanisms underlying this simple behaviour. However, the view that the amygdala is specialized only for fear conditioning is too narrow. Strong evidence supports an integral role for this region in other aversive states, such as anxiety, as well as in reward, providing a challenge for determining how neuronal activity within the amygdala could contribute to each of these distinct processes.

Cell-type- and projection-specific manipulations are providing a way forward. An example comes from the reverse translational study of anxiety described earlier. To probe the role of BLA–CeA projections in anxiety-related behaviour, the excitatory opsin, ChR2, was expressed in BLA neurons. Consistent with previous work, BLA-cell-body activation decreased time spent by mice in the open arms of an elevated plus maze (EPM), indicating an increase in anxiety-like behaviour<sup>36</sup>. However, on selective excitation of the BLA–CeL pathway an anxiolytic behavioural phenotype of exploration of the EPM open arms was observed. Thus, a glutamatergic BLA projection to the CeA can promote anxiolysis<sup>36</sup>.

In this study, the identity (that is, ON or OFF cell, PKC $\delta^+$  or PKC $\delta^-$ ) of the CeL neurons activated by BLA terminal optical stimulation was not determined. New evidence suggests that these neurons are CeL:PKC $\delta^+$ ; when ChR2 was selectively expressed in CeL:PKC $\delta^+$  neurons to allow their direct photoactivation, increased time in the open arms of the EPM was observed<sup>58</sup>. These data support the notion that a population of BLA neurons preferentially excites CeL:PKC $\delta^+$  neurons that lead to reduced anxiety, perhaps through the projection from CeL:PKC $\delta^+$  to CeM output neurons. Alternatively, the CeL:PKC $\delta^+$  neuronal projection to the BNST, also implicated in anxiety<sup>59,60</sup>, could mediate anxiolysis. These studies show that the BLA–CeA pathway does not universally promote aversive states. Of note, there are other pathways that originate in the BLA that have since been implicated in bidirectional regulation of anxiety. Optically driving BLA projections to the ventral HPC (vHPC) is anxiogenic<sup>61</sup>, whereas photostimulation of BLA inputs in the anterodorsal BNST (adBNST) is anxiolytic<sup>62</sup>. Photostimulation of the BLA–vHPC pathway also decreases social interaction<sup>63</sup>, an effect perhaps related to the increase in anxiety that is seen following BLA–vHPC activation<sup>64</sup>.

New findings continue to enlarge our conceptions of CeA circuitry in non-fear behaviour. In a recent study on feeding suppression, genetic targeting of CeL:PKC $\delta^+$  neurons was used to selectively express hM4Di to allow for reversible neuronal inhibition. In mice, reducing the activity of CeL:PKC $\delta^+$  neurons was found to block decreased feeding stimulated by sickness or unpalatable tastes<sup>58</sup>. Conversely, optogenetic excitation of CeL:PKC $\delta^+$  neurons dramatically decreased feeding in both hungry and sated mice, and this was not a confound of reduced activity or increased

anxiety. These results indicate that CeL:PKC $\delta^+$  neurons suppress feeding in response to anorexigenic stimuli<sup>58</sup>.

What inputs might normally serve to drive CeL:PKC $\delta^+$  neurons and suppress feeding? Tract tracing methods using cre-dependent retrograde and anterograde viral techniques can reveal the specific inputs and outputs of a given cell type<sup>65</sup>. Restriction of rabies virus to CeL:PKC $\delta^+$  neurons to allow labelling of presynaptic inputs, combined with c-fos immunohistochemistry to measure cellular activation, revealed multiple inputs to CeL:PKC $\delta^+$  neurons that are activated by anorexigenic signals, including the BLA, the parabrachial nucleus (PBN) and the insula (IN)<sup>58</sup>. These findings collectively demonstrate a unique pathway for feeding suppression that relies on CeL circuitry and that was not readily identified using pharmacology or lesion experimental approaches.

The identification of a role for amygdala circuitry in anxiolysis and feeding suppression makes one look at this circuitry differently in comparison to when only considering fear. How could these processes co-exist? Excitatory responses in the BLA to a fear cue activate CeL<sub>ON</sub>:PKC $\delta^-$  cells, some of which are SOM $^+$ , to initiate freezing. By contrast, excitatory responses in the BLA also activate CeL<sub>OFF</sub>:PKC $\delta^+$  cells to promote anxiolysis; to account for these findings, separate functional groups of BLA neurons projecting to CeL are proposed (Fig. 4). One route for the production of freezing and other measures of fear is through inhibition of CeL:PKC $\delta^+$  neurons that disinhibit CeM neurons; additional routes through CeL projections to other targets might also mediate conditioned freezing. Thus, the postsynaptic targets that mediate anxiolysis after excitation of CeL:PKC $\delta^+$  neurons may or may not be CeM neurons, and if they are, it is possible they are different CeM neurons to those mediating freezing.

In summary, recent studies have demonstrated that the BLA–CeA circuitry is involved in a diverse array of behaviours in addition to those related to fear. The key to understanding the production of these different behaviours is in the functional anatomy.

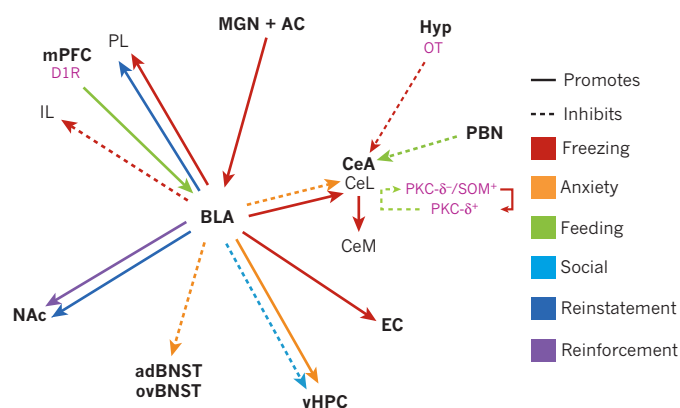
### Adding in reward

The above findings indicate that there must be a diversity of neuronal responses in the BLA, as opposed to only responses to fear cues; in fact, this was well known from *in vivo* electrophysiological recordings and is necessary to explain the effects of amygdala lesions on other behaviours besides fear conditioning. In parallel to early studies on fear conditioning, amygdala lesions were also found to impair reward-based behaviour<sup>66–72</sup>. For example, LA lesions prevent amphetamine place preference conditioning<sup>70</sup>, a procedure in which subjects learn to associate a particular spatial location with the reinforcing effects of the drug, and CeA lesions prevent conditioned orienting responses to reward-predictive cues<sup>68</sup>. Hence, the same lesions can impair Pavlovian conditioned behaviour in response to cues signalling either a rewarding or an aversive outcome.

Notably, amygdala lesions do not impair all behaviours emitted in response to reward-predictive cues, rather, they impair the ability to respond to cues in the face of changing reward value, leading to the hypothesis that learning mediated by the amygdala is related to the current, relative value of biologically significant outcomes<sup>8,73,74</sup>. This view is congruent with the broader notion that the amygdala provides information about the ‘state value’ of an organism, defined as the value of the overall situation of an organism at a given moment<sup>7</sup>.

The BLA and CeA seem to make distinct contributions to the representation of value, as revealed through procedures designed to change the current value of an outcome. BLA lesions impair the ability of value changes in specific reward outcomes to affect behaviour. Thus, the BLA is proposed to represent outcome value along with specific sensory features, allowing for discrimination among multiple outcomes of a similar valence<sup>23,72,73</sup>. By contrast, the CeA is considered to maintain a more general representation of the motivational significance of an outcome<sup>23,73</sup>.

Just as BLA neurons develop excitatory responses to a CS paired with aversive outcomes, they show excitatory responses to auditory, visual or olfactory CSs paired with rewarding outcomes, typically sweet liquid or food pellets<sup>75–79</sup>. As initially proposed for fear conditioning, evidence suggests that reward cue responses develop through long-lasting



**Figure 3 | Amygdalar circuits that are sufficient to alter behaviour in a diversity of domains.** Projection-specific effects as shown by optogenetic or pharmacogenetic manipulation. The solid or dotted lines indicate the promotion or inhibition of certain behaviours. The basolateral complex of the amygdala (BLA) encompasses the lateral and basal nuclei. Specific cell types are shown in pink. For simplicity, projections that are anatomically or electrophysiologically defined but have not been shown to have a causal relationship with behaviour are omitted. This is a selective picture of projections that have been directly manipulated, and is not meant to signify their importance over other anatomical connections. The actual connectivity of the amygdala with other brain regions is considerably more complex. AC, auditory cortex; adBNST, anterodorsal bed nucleus of the stria terminalis; CeA, central nucleus of the amygdala; CeL, lateral CeA; CeM, medial CeA; D1R, dopamine 1 receptor; EC, entorhinal cortex; Hyp, hypothalamus; IL, infralimbic; MGN, medial geniculate nucleus; mPFC, medial prefrontal cortex; NAc, nucleus accumbens; OT, oxytocin; ovBNST, oval nucleus of the BNST; PBN, parabrachial nucleus; PKC, protein kinase C; PL, prelimbic; SOM, somatostatin; vHPC, ventral hippocampus.

enhancement of glutamatergic inputs from sensory thalamus onto principal neurons in the LA<sup>78</sup>. Of note, the acquisition of conditioned responding to both fear and reward cues requires an NMDA-receptor-dependent increase in AMPA-receptor function in LA neurons<sup>44,46,78</sup>. The end result of this synaptic potentiation is that the cue is able to drive spiking in LA neurons, which in turn activates neuronal populations in downstream regions that contribute to the cue-triggered behaviour. Clearly, learning about environmental stimuli that predict the occurrence of food and other rewards is adaptive, as is learning about potential aversive outcomes. Thus, learning across a valence continuum of positive (rewarding) to negative (punishing or aversive) outcomes engages the amygdala.

### Amygdala neurons encode valence

Because both fear and reward cues recruit BLA neurons, these findings raise the question of how processing of fear and reward cues by amygdala networks is organized. For example, would neurons with excitatory responses to fear-predictive cues also show excitatory responses to reward-predictive cues, or is there segregation of positive- and negative-valenced signals? This question was addressed by directly comparing neural responses in the same subjects following training on both appetitive and aversive tasks. These within-subject comparisons in rodents and non-human primates consistently reveal populations of valence-selective neurons<sup>79–85</sup> such that some neurons excited by a fear cue do not respond to a reward cue, or show inhibition in the presence of the reward cue, and vice versa. By training subjects on two parallel cue outcome associations, with one cue followed by a rewarding outcome, and the other followed by an aversive outcome, and then reversing the cues assigned to each outcome, it was revealed that a substantial proportion of cue-selective neurons encode the outcome with which it is currently paired, not the sensory features of the cue itself<sup>81,82</sup>. Outcome-specific neuronal populations are consistent with valence-sensitive neuronal populations that are responsive to one valence, but not another. The finding that different BLA populations respond to a fear cue after learning than after extinction, when that cue signals no footshock<sup>86</sup>, can also be interpreted as valence



encoding. The observance of valence-specific neuronal activity is congruent with the notion that the amygdala is concerned with the relative value of the outcome, the occurrence of which is signalled by the cue.

Notably, valence-encoding is complemented by salience-encoding; some BLA units show excitatory responses to both fear and reward cues<sup>80,83,85</sup>, and these responses are correlated with measures of autonomic nervous system activation<sup>83</sup>. The salience of a stimulus is defined as the intensity of a stimulus and is a second dimension along which stimuli are encoded, in line with common models of emotion<sup>87</sup>. The salience responses may contribute to processes of arousal and attention that enhance processing within the amygdala or in target regions. This may be reflected behaviourally in better performance in real-time and in enhanced learning. A role in signalling the salience of stimuli is in agreement with the suggested contributions of the BLA in attention<sup>88,89</sup> and in enhancing memory storage in downstream regions<sup>90–92</sup>. The amygdala projects to basal forebrain cholinergic systems and midbrain dopaminergic systems, two means by which these effects may be mediated. For example, CeA projections to midbrain dopaminergic regions are required for conditioned orienting<sup>93</sup>. Furthermore, projections to sensory cortical regions, including primary visual cortex in primates<sup>21</sup>, may allow amygdala attentional signals to modulate stimulus processing<sup>94</sup>. In addition, findings suggest that attention modulates amygdala valence signals<sup>95</sup>.

What might lend a cell its functional phenotype? No obvious anatomical segregation of neurons that are sensitive to the reward or fear cue has been detected with electrophysiology<sup>83,96</sup>, demonstrating that neurons encoding two very different signals are intermingled within this structure. But wiring must still be fundamental to cell phenotype. A given BLA principal neuron, for example, is likely to be associated with reward or fear by virtue of its distinct connectivity, including both the pattern of specific inputs to that neuron and its projections to effector regions for expression of the adaptively appropriate behaviour. Amygdala neuronal pairs sensitive to stimuli of the same valence are more likely to show correlated neural activity than neuronal pairs sensitive to opposite valences, supporting the idea of functional networks<sup>96</sup>.

### Membership in a memory network

How do these functional networks for fear, reward or anxiety arise? Molecular genetic studies in mice have investigated the size, stability and required initial conditions for the formation of memory ensembles after fear conditioning. Using expression of the activity-dependent gene *Arc* to visualize activated LA neurons, the proportion of principal neurons participating in the fear memory trace is estimated to be 15–25%<sup>97,98</sup>. Although there may be a larger deterministic population, final membership in the trace can be biased by enhancing the activity level of individual LA neurons by multiple methods, including CREB overexpression<sup>97</sup> and brief optical activation immediately before tone–footshock pairing, which takes advantage of the temporal precision of optogenetics to prove that activity enhancement need only be present immediately before training<sup>98</sup>. Selective ablation after training of CREB-overexpressing neurons abolished conditioned freezing, providing strong evidence that these neurons are recruited into the memory trace<sup>99</sup>. Expressing the excitatory DREADD hM3Dq in a sparse population of LA neurons and activating these receptors during fear conditioning enhanced fear memory and biased the inclusion of these neurons into the fear trace. When the hM3Dq receptors are activated without presentation of the CS, the conditioned freezing response is partially recapitulated<sup>98</sup>. Importantly, similar mechanisms act for reward learning; the acquisition and maintenance of a cocaine-conditioned place preference depends on the recruitment of a small population of neurons in the LA<sup>100</sup>. These studies provide evidence that subsets of LA neurons stably participate in fear and reward memory traces, and there is similar evidence for ensembles in the BA<sup>101</sup>.

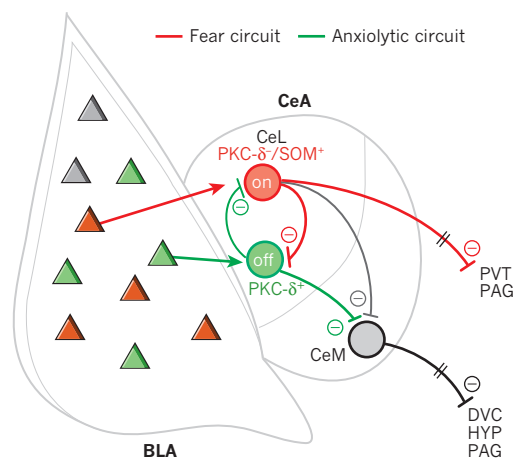
Although there do seem to be similarities in the cellular mechanisms that promote inclusion into memory networks, does that mean that any LA neuron could be either a ‘fear’ or ‘reward’ neuron, or are there constraints on ensemble membership? A recent study<sup>102</sup> provides evidence that valence-specific ensembles in the BLA (including the LA), once

formed, are immutable, which does not directly demonstrate that there are anatomical constraints on initial ensemble formation, but is congruent with that possibility. This study used ‘memory trace tagging’ approaches to determine whether a given neuron is limited to representing one valence or shows equipotentiality. ChR2 expression was limited to BLA or HPC neurons in the dentate gyrus (DG) that expressed c-fos on either fear (footshock) or reward (mate exposure) conditioning, thereby labelling a network of neurons that presumably constitute part of the memory trace. Later, optical activation of these networks supported either approach or avoidance following fear or reward conditioning, respectively, in a real-time place preference (RTPP) task, establishing the ‘valence’ of the labelled network. To examine the reversibility of valence assignment, a network of a given valence was activated during subsequent retraining with outcomes of the opposite valence. Whereas pairing an ensemble of DG neurons activated with one valence with the opposite valence US could produce a ‘switch’ in the valence of the tagged ensemble, this was not true for BLA ensembles<sup>102</sup>. It seems likely that this ‘fixed valence’ feature derives from the distinct wiring of BLA neurons into positive- and negative-valenced networks, and is in agreement with the observations made within the awake recording studies already discussed. Together, these findings raise the question of what the cellular identity and anatomical connectivity of those positive or negative valence ensembles might be.

Although reward and fear networks in the BLA seem to occupy an overlapping spatial location, a recent study using c-fos activation patterns paints a different picture for the CeA. By taking advantage of the different time courses for stimulation of c-fos mRNA and protein production, the CeA neuronal populations activated by two oppositely valenced stimuli were visualized using fluorescence immunohistochemistry and *in situ* hybridization within mice. Neurons activated by morphine were located primarily in the CeL, whereas neurons activated by footshock were mainly found within the CeM<sup>103</sup>. Although this study provides intriguing new information on the activation of CeA populations in a valence-specific manner, earlier studies already reviewed indicate that neurons within both the CeL and CeM contribute to diverse behaviours, so the functional implications of this result remain to be determined.

### Valence encoding in BLA inhibitory networks

An important key to understanding neural diversity is determining how individual neurons act within a circuit, and electrophysiology provides one means to query the role of a neuron. However, a drawback of *in vivo* extracellular recording studies is that the neuronal subtype, such as



**Figure 4 | Model of amygdala microcircuits that give rise to behaviour.**

New findings in the amygdala have updated our understanding of these microcircuits. Different populations of basolateral complex of the amygdala (BLA) neurons are proposed to activate distinct populations of lateral central nucleus of the amygdala (CeL) neurons to either promote fear or reduce anxiety. CeM, medial central nucleus of the amygdala; DVC, dorsal vagal complex; PAG, periaqueductal grey; PKC, protein kinase C; PVT, paraventricular nucleus of the thalamus; HYP, hypothalamus; SOM, somatostatin.

projection neuron or interneuron, can only be inferred from physiological measures. The ability to target opsin expression in defined neuronal subpopulations, in combination with electrophysiology, is overcoming this issue and has allowed us to gain new understanding of the inhibitory processes in the amygdala.

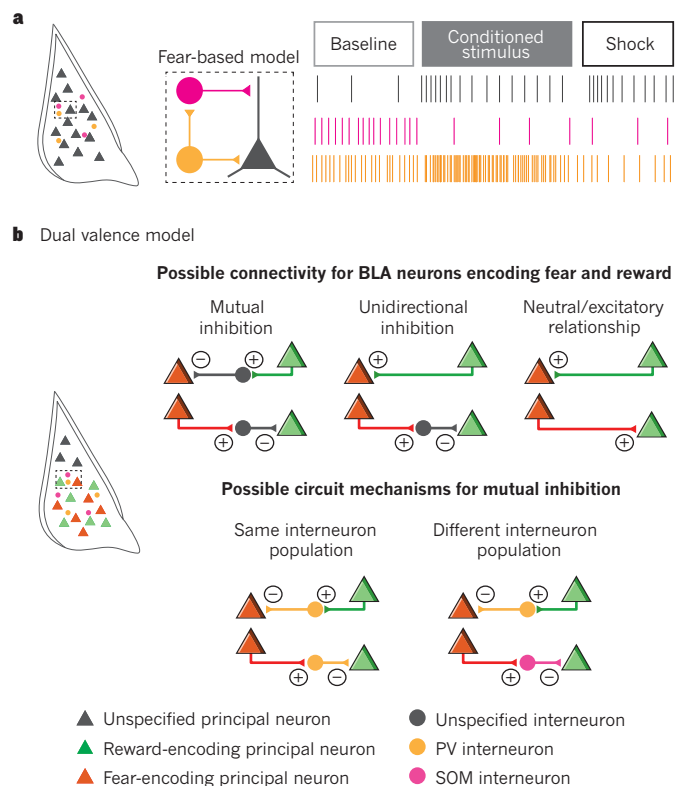
As described above, interactions among CeL inhibitory neurons, including inhibitory neurons that send projections outside the CeL itself, are crucial for the production of behavioural measures of fear. By contrast, in the BLA, inhibition entails suppression of excitatory principal projection neurons by local inhibitory interneurons and intercalated neurons. A recent study harnessed the power of optogenetics to study inhibitory BLA interneurons by temporally restricting excitation or inhibition of parvalbumin (PV) or SOM interneurons to the US footshock or the auditory CS, and observing the effect on fear conditioning<sup>104</sup>. Intriguing evidence of two different local circuit mechanisms that disinhibit principal neurons was found: during US presentation, PV neurons that provide perisomatic innervation of BLA principal neurons are inhibited by footshock, thereby directly disinhibiting BLA principal neurons; during the CS, PV neurons inhibit SOM interneurons that selectively contact principal neuron dendrites to disinhibit input-driven activity (Fig. 5a). Unit recordings were made during behaviour, and comparisons between these naturally occurring and optically evoked waveforms were used to convincingly ascribe neural correlates of behaviour to spike activity of PV and SOM interneurons. The electrophysiological activity of a subset of these interneurons during the CS and the US was the same as predicted by the model constructed from the optogenetic manipulations, providing support for the proposed local circuit mechanisms<sup>104</sup>. Thus, biologically significant stimuli and the cues that predict them change the firing of inhibitory interneurons, the activity of which gates the responsiveness of principal neurons.

Interestingly, the *in vivo* activity of PV and SOM interneurons is not limited to the role suggested by the above model. Of note, *in vivo* recordings found that a substantial subset of interneurons responded to the CS and US in an opposite way to that predicted by the model<sup>104</sup>, yet another example of neuronal response diversity. This is consistent with a view of parallel opposing circuits in the BLA, in which distinct populations of inhibitory neurons may govern the activity of BLA neurons mediating positive and negative valenced associations.

Together, the findings support proposed roles for inhibitory neurons in shaping functional networks in the amygdala<sup>6,12</sup>. Neuronal networks in the BLA may actively suppress other networks, the output of which would be incompatible with current behavioural requirements by feed forward inhibition onto a principal neuron through an intervening interneuron activated by a neighbouring principal neuron encoding a different valence or behaviour. To flexibly switch between different behavioural states, mutual inhibition among parallel competing networks is an attractive solution, although there may be other means of interaction among competing networks (Fig. 5). It is probable that long-range excitatory synapses, for example from the cortex, onto inhibitory interneurons (or GABAergic intercalated populations<sup>29</sup>) are crucial for the suppression of opposing networks<sup>105</sup> (for simplicity, external afferent influences are not depicted in Fig. 5b). Other means to inhibit opposing networks include changes in synaptic input; extinction, for example, induces heterosynaptic inhibition of thalamic inputs onto LA neurons<sup>105</sup>. In addition, the effect of interneurons on principal neurons is dynamic. New evidence indicates that inhibitory synapses onto principal cells change with experience; extinction training increases PV<sup>+</sup> perisomatic synaptic contacts on principal neurons that are activated during fear learning<sup>106</sup>, an example of how inhibition can remodel a network as the value of a cue changes.

### Valence, amygdala circuits and behaviour

Given its access to primary sensory information, the amygdala seems well suited to rapidly process and transmit information regarding the positive or negative valence of stimuli to bias behaviour in an adaptive manner. The evidence reviewed above suggests that networks within



**Figure 5 | Interneuron and principal neuron interactions within the basolateral complex of the amygdala (BLA).** **a**, Model of how interneurons expressing parvalbumin (PV) and somatostatin (SOM) interact with principal neurons to mediate fear conditioning<sup>89</sup>. **b**, The heterogeneity in PV interneuron responses is consistent with the diverse functionality of BLA principal neurons and raises the question of how BLA principal neurons may interact locally. Depicted are simplified scenarios for these interactions.

the amygdala are organized into distinct neural circuits for positive- and negative-valenced stimuli. How do these circuits come to affect behaviour?

To answer this question, we must map functionality onto the anatomical connections within the amygdala itself and onto projections from the amygdala. For fear conditioning, functional circuit mapping is relatively well developed, as described above. In the case of reward-related behaviours, locomotion, directed approach and manipulation of objects in the environment in order to obtain or interact with the reward are usually required. These types of coordinated, but flexible action patterns, engage corticostriatal circuitry. The BLA projects robustly to the NAc<sup>107</sup>, which mediates motivated responding to reward-predictive cues for both natural and drug rewards; hence the BLA–NAc projection could allow information regarding the current value of cues to affect reward-related behaviour, as suggested by earlier findings<sup>66,67,108</sup>. These ideas can be directly assessed using projection-specific manipulations during behaviour. The first findings along these lines reported that the BLA–NAc projection is in itself sufficient to support positive reinforcement, as demonstrated by intracranial self-stimulation<sup>109,110</sup> and RTPP, in which entry into a particular spatial location triggered stimulation of BLA–NAc terminals<sup>109</sup>. More recently, optical inhibition of the BLA–NAc projection (as well as the BLA–prelimbic (PL) projection) was found to prevent reinstatement of responding in reaction to a cocaine-paired cue in an animal model of relapse, further supporting a role for this projection in cue–reward associations<sup>111</sup>. These findings demonstrate that some BLA neurons involved in reward project to the NAc. Perhaps a subset of these receive efferents from the PFC; photoactivation of mPFC terminals in the BLA has been reported to increase instrumental responses for food<sup>112</sup>. Collectively these studies have begun to define amygdala circuits that contribute



to reward-related behaviour, although the circuit analysis is not as far along for reward as it is for fear.

As function continues to be mapped onto amygdala circuitry, our understanding of how neural signals in the amygdala affect behaviour will continue to grow. A wide range of behavioural changes has been achieved by optogenetic or pharmacogenetic manipulation of the projections studied so far (Fig. 3). Two observations can be made: first, multiple projections leading from the amygdala affect a single behaviour, and second, behaviours of different or opposite valence are affected by projections between the same two brain regions. Both conditioned freezing and anxiety exemplify the first observation. In the case of anxiety, for example, the finding that three different BLA efferents influence anxiety-like behaviour<sup>36,61,62</sup> suggests that even at the level of the comparatively simple system of the amygdala, neural control of a single behaviour is not reduced to one serial pathway but is normally accomplished by multiple circuits, although these circuits potentially interact. The second observation that the same pathway may affect very different behaviours is exemplified by the role of the BLA–PL pathway in both fear and relapse to reward seeking. In each of these behavioural procedures<sup>111,113</sup>, photoinhibition of the pathway impaired the behavioural effects of the conditioned cue; this pathway may be valence-independent, or may carry mixed fibres. These possibilities can be distinguished using the types of approaches discussed here.

This new functional map supports a view of the amygdala as a composite of parallel circuits that contribute to multiple behavioural states. Although there may be a substantial degree of overlap among the circuit elements, the circuits are differentiated by the specific details of their neural connections, both afferent and efferent, and the patterns of activation of those connections. Hence, to fully understand the relationship between neurons that transmit information about the valence of a stimulus and subsequent behaviour, neural circuit analysis is key. This way of thinking about the amygdala is different from past conceptions of it as a fear hub or as a circuit providing a readout of positive or negative affect in simple terms. Instead, the emphasis is on understanding the behaviourally relevant functions of paths of information flow through these regions, including how diverse, primarily sensory, inputs might interact locally to produce varied downstream functional effects.

Although we make the case that the amygdala contributes to a diverse set of behaviours, as revealed by the fine-grained analysis of circuits, the notion of valence remains a useful heuristic. Sensitivity to the valence of a stimulus, of whether it is good or bad, is crucially important to ensure appropriate behavioural responses that promote approach and the acquisition of food, safety and social partners, and in alternative circumstances, that promote vigilance, avoidance or aggression towards threats. Each of the behaviours affected by amygdala manipulations falls somewhere along the continuum of 'good' to 'bad'. We note that organizing adaptive behaviours only along the single dimension of valence is certainly an oversimplification, but this may capture an essential feature of these diverse behaviours that engages the amygdala.

## Looking forward

Recently, we have made great strides in delineating the functional microcircuitry of the amygdala using the new technologies of optogenetics, pharmacogenetics and viral-based tract tracing that take advantage of gene-targeting. Crucially, these approaches have been used in partnership with state-of-the-art electrophysiology and careful behavioural analysis. This programme of circuit analysis is equally applicable to the study of other neural systems and is a way forward towards deeper knowledge of the functions that emerge from neural circuits.

As we look to the future of research on amygdala circuits, we consider areas that deserve attention. It is important to define amygdala functional microcircuitry in preclinical models of human behavioural disorders. Animal and human studies implicate the amygdala

in anxiety disorders, autism and addiction. In the case of addiction, chronic alcohol use alters neural transmission in the CeA, and these changes have been linked to excessive alcohol use<sup>114</sup>. Furthermore, alcohol and drug seeking triggered by auditory cues requires the BLA<sup>115,116</sup>, whereas memory traces related to the smell and taste of alcohol that drive relapse are stored in the CeA<sup>117</sup>. In the case of autism, understanding the mediation of social interaction by the amygdala is highly relevant<sup>64,118</sup>, adding momentum to this line of research<sup>63</sup>.

Of great interest is determining how information from the amygdala affects downstream cortical circuits. Much research indicates that neural connectivity with the orbital prefrontal cortex is important for updating cue values after changes in their associated outcome<sup>8,119,120</sup>. In addition to the orbital prefrontal cortex, BLA neurons project strongly to the medial PFC. The function of BLA signals in these cortical regions is less likely to be closely tied to discrete behavioural output. Instead, these BLA–cortical projections are proposed to mediate the impact of Pavlovian associations on decision making<sup>7,121</sup>. In addition, electrophysiological recordings have uncovered dynamic experience-dependent interactions between the amygdala, PFC and HPC in the entrainment of oscillations of different frequencies<sup>122,123</sup>, but their causal role in behaviour is not clear. Selective projection manipulations may reveal the circuit and behaviour impacts of these long-range interactions. New studies manipulating specific amygdala projections to the cortex are beginning to address these issues<sup>111,113</sup>.

Increasingly, the model systems available for circuit analysis will include primates as genetically accessible non-human primate models are developed and viral-mediated tools for neural manipulation are optimized, allowing for further investigation of species similarities and differences in amygdala function. Moving forward, it is crucial to complement causal manipulations with measurements of neural activity during a wide variety of behaviours across multiple species.

Although we champion the ability to manipulate circuit components that can be isolated with genetic or anatomical features, the existing tools still have limitations that prevent a comprehensive understanding of these circuits. Genetically encodable tools for neural manipulation allow far greater targeting specificity than before, but it is unlikely that all of the neurons that project from one region to another, or that share a genetic marker, have identical functions. Thus, we may still be observing a 'majority vote' for a given behavioural readout when manipulating any circuit component, and only more specific targeting strategies will reveal the functional minority populations. Along these lines, the synchrony and timing of most photostimulation experiments are not physiological and could disturb important rhythmic interactions across distal networks in ways we do not understand. To solve these issues, we need a greater library of molecular markers and tools to target combinatorial expression patterns, and we need the means to induce more naturalistic activity patterns in neurons. Basic science insights into molecules, synapses, cells and circuits will need to be synthesized to achieve this level of understanding. ■

Received 18 August; accepted 3 December 2014.

- McDonald, A. J. Cortical pathways to the mammalian amygdala. *Prog. Neurobiol.* **55**, 257–332 (1998).
- Jarvis, E. D. et al. Avian brains and a new understanding of vertebrate brain evolution. *Nature Rev. Neurosci.* **6**, 151–159 (2005).
- Johnston, J. B. Further contributions to the study of the evolution of the forebrain. *J. Comp. Neurol.* **35**, 337–481 (1923).
- Kappers, C. U. A., Huber, G. C. & Crosby, E. C. *The Comparative Anatomy of the Nervous System of Vertebrates, Including Man* (Macmillan, 1936).
- Lanuza, E., Belekova, M., Martínez-Marcos, A., Font, C. & Martínez-García, F. Identification of the reptilian basolateral amygdala: an anatomical investigation of the afferents to the posterior dorsal ventricular ridge of the lizard *Podarcis hispanica*. *Eur. J. Neurosci.* **10**, 3517–3534 (1998).
- Duvarci, S. & Pare, D. Amygdala microcircuits controlling learned fear. *Neuron* **82**, 966–980 (2014).
- Morrison, S. E. & Salzman, C. D. Re-valuing the amygdala. *Curr. Opin. Neurobiol.* **20**, 221–230 (2010).

8. Murray, E. A. The amygdala, reward and emotion. *Trends Cogn. Sci.* **11**, 489–497 (2007).
9. Stamatakis, A. M. et al. Amygdala and bed nucleus of the stria terminalis circuitry: implications for addiction-related behaviors. *Neuropharmacology* **76**, 320–328 (2014).
10. Johansen, J. P., Cain, C. K., Ostroff, L. E. & LeDoux, J. E. Molecular mechanisms of fear learning and memory. *Cell* **147**, 509–524 (2011).
11. Pape, H.-C. & Pare, D. Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. *Physiol. Rev.* **90**, 419–463 (2010).
12. Ehrlich, I. et al. Amygdala inhibitory circuits and the control of fear memory. *Neuron* **62**, 757–771 (2009).
13. Brown, S. & Schäfer, E. An investigation into the functions of the occipital and temporal lobes of the monkey's brain. *Phil. Trans. R. Soc. B* **179**, 303–327 (1888).
14. Klüver, H. & Bucy, P. 'Psychic blindness' and other symptoms following bilateral temporal lobectomy in Rhesus monkeys. *Am. J. Physiol.* **119**, 352–353 (1937).
15. Weiskrantz, L. Behavioral changes associated with ablation of the amygdaloid complex in monkeys. *J. Comp. Physiol. Psychol.* **49**, 381–391 (1956).
16. LeDoux, J. E., Cicchetti, P., Xagoraris, A. & Romanski, L. M. The lateral amygdaloid nucleus: sensory interface of the amygdala in fear conditioning. *J. Neurosci.* **10**, 1062–1069 (1990).
17. Blanchard, D. C. & Blanchard, R. J. Innate and conditioned reactions to threat in rats with amygdaloid lesions. *J. Comp. Physiol. Psychol.* **81**, 281–290 (1972).
18. Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **372**, 669–672 (1994).
19. Anderson, A. K. & Phelps, E. A. Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature* **411**, 305–309 (2001).
20. Marowsky, A., Yanagawa, Y., Obata, K. & Vogt, K. E. A specialized subclass of interneurons mediates dopaminergic facilitation of amygdala function. *Neuron* **48**, 1025–1037 (2005).
21. Freese, J. L. & Amaral, D. G. The organization of projections from the amygdala to visual cortical areas TE and V1 in the macaque monkey. *J. Comp. Neurol.* **486**, 295–317 (2005).
22. Chareyron, L. J., Banta Lavenex, P., Amaral, D. G. & Lavenex, P. Stereological analysis of the rat and monkey amygdala. *J. Comp. Neurol.* **519**, 3218–3239 (2011).
23. Corbit, L. H. & Balleine, B. W. Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of pavlovian-instrumental transfer. *J. Neurosci.* **25**, 962–970 (2005).
24. Holland, P. C. & Gallagher, M. Double dissociation of the effects of lesions of basolateral and central amygdala on conditioned stimulus-potentiated feeding and Pavlovian-instrumental transfer. *Eur. J. Neurosci.* **17**, 1680–1694 (2003).
25. Quirk, G. J., Armony, J. L. & LeDoux, J. E. Fear conditioning enhances different temporal components of tone-evoked spike trains in auditory cortex and lateral amygdala. *Neuron* **19**, 613–624 (1997).
26. Quirk, G. J., Repa, C. & LeDoux, J. E. Fear conditioning enhances short-latency auditory responses of lateral amygdala neurons: parallel recordings in the freely behaving rat. *Neuron* **15**, 1029–1039 (1995).
- This is a seminal study showing the increased responding of LA neurons to a CS after fear conditioning.**
27. LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E. & Phelps, E. A. Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron* **20**, 937–945 (1998).
28. Morris, J. S., Ohman, A. & Dolan, R. J. Conscious and unconscious emotional learning in the human amygdala. *Nature* **393**, 467–470 (1998).
29. Amano, T., Unal, C. T. & Paré, D. Synaptic correlates of fear extinction in the amygdala. *Nature Neurosci.* **13**, 489–494 (2010).
30. Milad, M. R. & Quirk, G. J. Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature* **420**, 70–74 (2002).
31. Phelps, E. A., Delgado, M. R., Nearing, K. I. & LeDoux, J. E. Extinction learning in humans: role of the amygdala and vmPFC. *Neuron* **43**, 897–905 (2004).
32. Nader, K., Schafe, G. E. & LeDoux, J. E. Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* **406**, 722–726 (2000).
33. Monfils, M.-H., Cowansage, K. K., Klann, E. & LeDoux, J. E. Extinction-reconsolidation boundaries: key to persistent attenuation of fear memories. *Science* **324**, 951–955 (2009).
34. Schiller, D. et al. Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* **463**, 49–53 (2010).
35. Etkin, A., Prater, K. E., Schatzberg, A. F., Menon, V. & Greicius, M. D. Disrupted amygdalar subregion functional connectivity and evidence of a compensatory network in generalized anxiety disorder. *Arch. Gen. Psychiatry* **66**, 1361–1372 (2009).
36. Tye, K. M. et al. Amygdala circuitry mediating reversible and bidirectional control of anxiety. *Nature* **471**, 358–362 (2011).
- This was the first study to use optogenetic projection-specific manipulations; it showed that activation or inhibition of BLA projections to the CeL nucleus could cause anxiolytic or anxiogenic effects on behaviour, respectively.**
37. Davis, M. The role of the amygdala in fear and anxiety. *Annu. Rev. Neurosci.* **15**, 353–375 (1992).
38. LeDoux, J. E. Emotion circuits in the brain. *Annu. Rev. Neurosci.* **23**, 155–184 (2000).
39. Nader, K., Majidshad, P., Amorapanth, P. & LeDoux, J. E. Damage to the lateral and central, but not other, amygdaloid nuclei prevents the acquisition of auditory fear conditioning. *Learn. Mem.* **8**, 156–163 (2001).
40. Collins, D. R. & Paré, D. Differential fear conditioning induces reciprocal changes in the sensory responses of lateral amygdala neurons to the CS<sup>+</sup> and CS. *Learn. Mem.* **7**, 97–103 (2000).
41. Maren, S. Auditory fear conditioning increases CS-elicited spike firing in lateral amygdala neurons even after extensive overtraining. *Eur. J. Neurosci.* **12**, 4047–4054 (2000).
42. Rogan, M. T., Stäubli, U. V. & LeDoux, J. E. Fear conditioning induces associative long-term potentiation in the amygdala. *Nature* **390**, 604–607 (1997).
43. McKernan, M. G. & Shinnick-Gallagher, P. Fear conditioning induces a lasting potentiation of synaptic currents *in vitro*. *Nature* **390**, 607–611 (1997).
- Along with ref. 42, this was the first evidence to show synaptic enhancement onto LA neurons after fear conditioning.**
44. Clem, R. L. & Huganir, R. L. Calcium-permeable AMPA receptor dynamics mediate fear memory erasure. *Science* **330**, 1108–1112 (2010).
45. Rumpel, S., LeDoux, J., Zador, A. & Malinow, R. Postsynaptic receptor trafficking underlying a form of associative learning. *Science* **308**, 83–88 (2005).
46. Johansen, J. P. et al. Optical activation of lateral amygdala pyramidal cells instructs associative fear learning. *Proc. Natl Acad. Sci. USA* **107**, 12692–12697 (2010).
47. Nabavi, S. et al. Engineering a memory with LTD and LTP. *Nature* **511**, 348–352 (2014).
48. Kapp, B. S., Frysinger, R. C., Gallagher, M. & Haselton, J. R. Amygdala central nucleus lesions: effect on heart rate conditioning in the rabbit. *Physiol. Behav.* **23**, 1109–1117 (1979).
49. Hitchcock, J. & Davis, M. Lesions of the amygdala, but not of the cerebellum or red nucleus, block conditioned fear as measured with the potentiated startle paradigm. *Behav. Neurosci.* **100**, 11–22 (1986).
50. LeDoux, J. E., Iwata, J., Cicchetti, P. & Reis, D. J. Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. *J. Neurosci.* **8**, 2517–2529 (1988).
51. Viviani, D. et al. Oxytocin selectively gates fear responses through distinct outputs from the central amygdala. *Science* **333**, 104–107 (2011).
52. Cioocchi, S. et al. Encoding of conditioned fear in central amygdala inhibitory circuits. *Nature* **468**, 277–282 (2010).
53. Haubensak, W. et al. Genetic dissection of an amygdala microcircuit that gates conditioned fear. *Nature* **468**, 270–276 (2010).
- Together with ref. 52 this study identified functionally and genetically distinct populations of neurons in the CeL in the expression of conditioned fear.**
54. Li, H. et al. Experience-dependent modification of a central amygdala fear circuit. *Nature Neurosci.* **16**, 332–339 (2013).
- This article reports that experience-dependent plasticity occurs at LA–CeL:SOM<sup>+</sup> synapses, demonstrating that amygdala plasticity occurs in more than just the LA.**
55. Penzo, M. A., Robert, V. & Li, B. Fear conditioning potentiates synaptic transmission onto long-range projection neurons in the lateral subdivision of central amygdala. *J. Neurosci.* **34**, 2432–2437 (2014).
56. Sparta, D. R. et al. Inhibition of projections from the basolateral amygdala to the entorhinal cortex disrupts the acquisition of contextual fear. *Front. Behav. Neurosci.* **8**, 129 (2014).
57. Knobloch, H. S. et al. Evoked axonal oxytocin release in the central amygdala attenuates fear response. *Neuron* **73**, 553–566 (2012).
58. Cai, H., Haubensak, W., Anthony, T. E. & Anderson, D. J. Central amygdala PKC $\delta$  neurons mediate the influence of multiple anorexigenic signals. *Nature Neurosci.* **17**, 1240–1248 (2014).
- This study showed that PKC $\delta$  neurons suppress feeding and are anxiolytic, and using a 'cre-out' strategy demonstrated opposing functions for PKC $\delta$  and PKC $\alpha$  neurons.**
59. Jennings, J. H. et al. Distinct extended amygdala circuits for divergent motivational states. *Nature* **496**, 224–228 (2013).
60. Kim, S.-Y. et al. Diverging neural pathways assemble a behavioural state from separable features in anxiety. *Nature* **496**, 219–223 (2013).
61. Felix-Ortiz, A. C. et al. BLA to vHPC inputs modulate anxiety-related behaviors. *Neuron* **79**, 658–664 (2013).
62. Kim, S.-Y. et al. Diverging neural pathways assemble a behavioural state from separable features in anxiety. *Nature* **496**, 219–223 (2013).
63. Felix-Ortiz, A. C. & Tye, K. M. Amygdala inputs to the ventral hippocampus bidirectionally modulate social behavior. *J. Neurosci.* **34**, 586–595 (2014).
64. Allsop, S. A., Vander Weele, C. M., Wichmann, R. & Tye, K. M. Optogenetic insights on the relationship between anxiety-related behaviors and social deficits. *Front. Behav. Neurosci.* **8**, 241 (2014).
65. Wall, N. R., Wickersham, I. R., Cetin, A., De La Parra, M. & Callaway, E. M. Monosynaptic circuit tracing *in vivo* through Cre-dependent targeting and complementation of modified rabies virus. *Proc. Natl Acad. Sci. USA* **107**, 21848–21853 (2010).
66. Cador, M., Robbins, T. W. & Everitt, B. J. Involvement of the amygdala in stimulus-reward associations: interaction with the ventral striatum. *Neuroscience* **30**, 77–86 (1989).
67. Everitt, B. J., Cador, M. & Robbins, T. W. Interactions between the amygdala and ventral striatum in stimulus-reward associations: studies using a second-order schedule of sexual reinforcement. *Neuroscience* **30**, 63–75 (1989).
- This study, along with ref. 66, provided early evidence that amygdala projections to the NAc mediate the effects of Pavlovian stimuli predictive of reward on behaviour.**
68. Gallagher, M., Graham, P. W. & Holland, P. C. The amygdala central nucleus and appetitive Pavlovian conditioning: lesions impair one class of conditioned behavior. *J. Neurosci.* **10**, 1906–1911 (1990).
69. Hatfield, T., Han, J. S., Conley, M., Gallagher, M. & Holland, P. Neurotoxic lesions



- of basolateral, but not central, amygdala interfere with Pavlovian second-order conditioning and reinforcer devaluation effects. *J. Neurosci.* **16**, 5256–5265 (1996).
70. Hiroi, N. & White, N. M. The lateral nucleus of the amygdala mediates expression of the amphetamine-produced conditioned place preference. *J. Neurosci.* **11**, 2107–2116 (1991).
  71. McDonald, R. J. & White, N. M. A triple dissociation of memory systems: hippocampus, amygdala, and dorsal striatum. *Behav. Neurosci.* **107**, 3–22 (1993).
  72. Málková, L., Gaffan, D. & Murray, E. A. Excitotoxic lesions of the amygdala fail to produce impairment in visual learning for auditory secondary reinforcement but interfere with reinforcer devaluation effects in rhesus monkeys. *J. Neurosci.* **17**, 6011–6020 (1997).
  73. Balleine, B. W. & Killcross, S. Parallel incentive processing: an integrated view of amygdala function. *Trends Neurosci.* **29**, 272–279 (2006).
  74. Baxter, M. G. & Murray, E. A. The amygdala and reward. *Nature Rev. Neurosci.* **3**, 563–573 (2002).
  75. Sanghera, M. K., Rolls, E. T. & Roper-Hall, A. Visual responses of neurons in the dorsolateral amygdala of the alert monkey. *Exp. Neurol.* **63**, 610–626 (1979).
  76. Schoenbaum, G., Chiba, A. A. & Gallagher, M. Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neurosci.* **1**, 155–159 (1998).
  77. Tye, K. M. & Janak, P. H. Amygdala neurons differentially encode motivation and reinforcement. *J. Neurosci.* **27**, 3937–3945 (2007).
  78. Tye, K. M., Stuber, G. D., de Ridder, B., Bonci, A. & Janak, P. H. Rapid strengthening of thalamo-amygdala synapses mediates cue-reward learning. *Nature* **453**, 1253–1257 (2008).
  - This study demonstrated a causal relationship between synaptic potentiation in the amygdala and cue-reward learning, and showed amygdala neurons increase responses *in vivo* with cue-reward learning.**
  79. Uwano, T., Nishijo, H., Ono, T. & Tamura, R. Neuronal responsiveness to various sensory stimuli, and associative learning in the rat amygdala. *Neuroscience* **68**, 339–361 (1995).
  80. Belova, M. A., Paton, J. J., Morrison, S. E. & Salzman, C. D. Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron* **55**, 970–984 (2007).
  81. Paton, J. J., Belova, M. A., Morrison, S. E. & Salzman, C. D. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* **439**, 865–870 (2006).
  - In this study, electrophysiological recordings showed that different populations of primate amygdala neurons encoded visual stimuli that predicted positive or negative outcomes.**
  82. Schoenbaum, G., Chiba, A. A. & Gallagher, M. Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. *J. Neurosci.* **19**, 1876–1884 (1999).
  - This was the first electrophysiological recording study demonstrating the ability of amygdala neurons to track changing outcomes across a reversal task.**
  83. Shabel, S. J. & Janak, P. H. Substantial similarity in amygdala neuronal activity during conditioned appetitive and aversive emotional arousal. *Proc. Natl Acad. Sci. USA* **106**, 15031–15036 (2009).
  - This study suggested that populations of amygdala neurons that encoded positive and negative outcomes were only partially non-overlapping; the overlapping population may encode salience.**
  84. Shabel, S. J., Schairer, W., Donahue, R. J., Powell, V. & Janak, P. H. Similar neural activity during fear and disgust in the rat basolateral amygdala. *PLoS ONE* **6**, e27797 (2011).
  85. Sangha, S., Chadick, J. Z. & Janak, P. H. Safety encoding in the basal amygdala. *J. Neurosci.* **33**, 3744–3751 (2013).
  86. Herry, C. *et al.* Switching on and off fear by distinct neuronal circuits. *Nature* **454**, 600–606 (2008).
  87. Russell, J. A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **39**, 1161–1178 (1980).
  88. Holland, P. C. & Gallagher, M. Amygdala circuitry in attentional and representational processes. *Trends Cogn. Sci.* **3**, 65–73 (1999).
  89. Roesch, M. R., Esber, G. R., Li, J., Daw, N. D. & Schoenbaum, G. Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. *Eur. J. Neurosci.* **35**, 1190–1200 (2012).
  90. McGaugh, J. L. The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annu. Rev. Neurosci.* **27**, 1–28 (2004).
  91. Huff, M. L., Miller, R. L., Deisseroth, K., Moorman, D. E. & LaLumiere, R. T. Posttraining optogenetic manipulations of basolateral amygdala activity modulate consolidation of inhibitory avoidance memory in rats. *Proc. Natl Acad. Sci. USA* **110**, 3597–3602 (2013).
  92. Popescu, A. T., Saghyian, A. A. & Paré, D. NMDA-dependent facilitation of corticostriatal plasticity by the amygdala. *Proc. Natl Acad. Sci. USA* **104**, 341–346 (2007).
  93. Han, J. S., McMahan, R. W., Holland, P. & Gallagher, M. The role of an amygdalo-nigrostriatal pathway in associative learning. *J. Neurosci.* **17**, 3913–3919 (1997).
  94. Vuilleumier, P., Richardson, M. P., Armony, J. L., Driver, J. & Dolan, R. J. Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nature Neurosci.* **7**, 1271–1278 (2004).
  95. Peck, C. J. & Salzman, C. D. Amygdala neural activity reflects spatial attention towards stimuli promising reward or threatening punishment. *eLife* **3**, e04478 (2014).
  96. Zhang, W. *et al.* Functional circuits and anatomical distribution of response properties in the primate amygdala. *J. Neurosci.* **33**, 722–733 (2013).
  97. Han, J.-H. *et al.* Neuronal competition and selection during memory formation. *Science* **316**, 457–460 (2007).
  98. Yiu, A. P. *et al.* Neurons are recruited to a memory trace based on relative neuronal excitability immediately before training. *Neuron* **83**, 722–735 (2014).
  99. Han, J.-H. *et al.* Selective erasure of a fear memory. *Science* **323**, 1492–1496 (2009).
  - This study provided causal evidence for a stable fear memory engram in the LA by ablating a small proportion of LA neurons overexpressing CREB.**
  100. Hsiang, H.-L. L. *et al.* Manipulating a ‘cocaine engram’ in mice. *J. Neurosci.* **34**, 14115–14127 (2014).
  101. Reijmers, L. G., Perkins, B. L., Matsuo, N. & Mayford, M. Localization of a stable neural correlate of associative memory. *Science* **317**, 1230–1233 (2007).
  102. Redondo, R. L. *et al.* Bidirectional switch of the valence associated with a hippocampal contextual memory engram. *Nature* **513**, 426–430 (2014).
  - This study used neuronal tagging to express ChR2 in valence-specific networks, demonstrating that positive and negative valenced networks in the BLA cannot be reversed to the opposite valence by retraining.**
  103. Xiu, J. *et al.* Visualizing an emotional valence map in the limbic forebrain by TAI-FISH. *Nature Neurosci.* **17**, 1552–1559 (2014).
  104. Wolff, S. B. E. *et al.* Amygdala interneuron subtypes control fear learning through disinhibition. *Nature* **509**, 453–458 (2014).
  - Demonstration of unique roles for PV<sup>+</sup> and SOM<sup>+</sup> interneurons in combination with *in vivo* electrophysiology in behaving mice to provide new evidence for inhibitory networks contributing to fear conditioning.**
  105. Cho, J.-H., Deisseroth, K. & Bolshakov, V. Y. Synaptic encoding of fear extinction in mPFC-amygdala circuits. *Neuron* **80**, 1491–1507 (2013).
  106. Trouche, S., Sasaki, J. M., Tu, T. & Reijmers, L. G. Fear extinction causes target-specific remodeling of perisomatic inhibitory synapses. *Neuron* **80**, 1054–1065 (2013).
  107. Kelley, A. E., Domesick, V. B. & Nauta, W. J. The amygdalo-striatal projection in the rat—an anatomical study by anterograde and retrograde tracing methods. *Neuroscience* **7**, 615–630 (1982).
  108. Ambroggi, F., Ishikawa, A., Fields, H. L. & Nicola, S. M. Basolateral amygdala neurons facilitate reward-seeking behavior by exciting nucleus accumbens neurons. *Neuron* **59**, 648–661 (2008).
  109. Britt, J. P. *et al.* Synaptic and behavioral profile of multiple glutamatergic inputs to the nucleus accumbens. *Neuron* **76**, 790–803 (2012).
  110. Stuber, G. D. *et al.* Excitatory transmission from the amygdala to nucleus accumbens facilitates reward seeking. *Nature* **475**, 377–380 (2011).
  111. Stefanik, M. T. & Kalivas, P. W. Optogenetic dissection of basolateral amygdala projections during cue-induced reinstatement of cocaine seeking. *Front. Behav. Neurosci.* **7**, 213 (2013).
  112. Land, B. B. *et al.* Medial prefrontal D1 dopamine neurons control food intake. *Nature Neurosci.* **17**, 248–253 (2014).
  113. Senn, V. *et al.* Long-range connectivity defines behavioral specificity of amygdala neurons. *Neuron* **81**, 428–437 (2014).
  114. Roberto, M., Gilpin, N. W. & Siggins, G. R. The central amygdala and alcohol: role of  $\gamma$ -aminobutyric acid, glutamate, and neuropeptides. *Cold Spring Harb. Perspect. Med.* **2**, a012195 (2012).
  115. Buffalari, D. M. & See, R. E. Amygdala mechanisms of Pavlovian psychostimulant conditioning and relapse. *Curr. Top. Behav. Neurosci.* **3**, 73–99 (2010).
  116. Chaudhri, N., Woods, C. A., Sahuque, L. L., Gill, T. M. & Janak, P. H. Unilateral inactivation of the basolateral amygdala attenuates context-induced renewal of Pavlovian-conditioned alcohol-seeking. *Eur. J. Neurosci.* **38**, 2751–2761 (2013).
  117. Barak, S. *et al.* Disruption of alcohol-related memories by mTORC1 inhibition prevents relapse. *Nature Neurosci.* **16**, 1111–1117 (2013).
  118. Baron-Cohen, S. *et al.* The amygdala theory of autism. *Neurosci. Biobehav. Rev.* **24**, 355–364 (2000).
  119. Saddoris, M. P., Gallagher, M. & Schoenbaum, G. Rapid associative encoding in basolateral amygdala depends on connections with orbitofrontal cortex. *Neuron* **46**, 321–331 (2005).
  120. Morrison, S. E., Saez, A., Lau, B. & Salzman, C. D. Different time courses for learning-related changes in amygdala and orbitofrontal cortex. *Neuron* **71**, 1127–1140 (2011).
  121. Seymour, B. & Dolan, R. Emotion, decision making, and the amygdala. *Neuron* **58**, 662–671 (2008).
  122. Likhtik, E., Stuijens, J. M., Topiwala, M. A., Harris, A. Z. & Gordon, J. A. Prefrontal entrainment of amygdala activity signals safety in learned fear and innate anxiety. *Nature Neurosci.* **17**, 106–113 (2014).
  123. Seidenbecher, T., Laxmi, T. R., Stork, O. & Pape, H.-C. Amygdalar and hippocampal theta rhythm synchronization during fear memory retrieval. *Science* **301**, 846–850 (2003).

**Acknowledgements** P.H.J. acknowledges funding from US National Institutes of Health grants DA015096, AA014925, AA17072. K.M.T. is a New York Stem Cell Foundation–Robertson Investigator and acknowledges funding from the JPB Foundation, PIIF, PNDRF, NARSAD Young Investigator Award, Whitehead Career Development Chair, and NIH grant MH102441. We thank K. Vitale for input regarding interneurons and network selection, G. Calhoun and P. Namburi for input on Fig. 5, R. Keiflin for assistance with Fig. 3, B. Saunders for comments on our text, J. Gabrieli for input on human amygdala research, I. Choi for assistance illustrating Fig. 1 and all the members of our laboratories for valuable discussion.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at [go.nature.com/v7haxd](http://go.nature.com/v7haxd). Correspondence should be addressed to P.H.J. ([patricia.janak@jhu.edu](mailto:patricia.janak@jhu.edu)) or K.M.T. ([kaytye@mit.edu](mailto:kaytye@mit.edu)).



# The biology of innate lymphoid cells

David Artis<sup>1</sup> & Hergen Spits<sup>2</sup>

**The innate immune system is composed of a diverse array of evolutionarily ancient haematopoietic cell types, including dendritic cells, monocytes, macrophages and granulocytes. These cell populations collaborate with each other, with the adaptive immune system and with non-haematopoietic cells to promote immunity, inflammation and tissue repair. Innate lymphoid cells are the most recently identified constituents of the innate immune system and have been the focus of intense investigation over the past five years. We summarize the studies that formally identified innate lymphoid cells and highlight their emerging roles in controlling tissue homeostasis in the context of infection, chronic inflammation, metabolic disease and cancer.**

**T**he innate and adaptive immune systems have evolved to simultaneously facilitate peaceful cohabitation with the trillions of beneficial microorganisms that constitute the microbiota, to provide host defence against infectious agents, and to initiate the repair and remodelling processes that restore and maintain tissue homeostasis<sup>1,2</sup>. Groundbreaking studies over the past five years formally identified innate lymphoid cells (ILCs) as part of the innate immune system; these cells can directly communicate with a wide variety of haematopoietic and non-haematopoietic cells to orchestrate immunity, inflammation and homeostasis in multiple tissues throughout the body<sup>3–5</sup>.

ILCs are a distinct arm of the innate immune system that are regulated by multiple endogenous mammalian cell-derived factors including neuropeptides, hormones, eicosanoids, cytokines and other alarmins<sup>3–5</sup>. The specialized distribution of ILCs in lymphoid and non-lymphoid tissues across multiple species, coupled with their functional heterogeneity<sup>3</sup>, has provoked a fundamental reassessment of how ILCs integrate innate and adaptive immune responses and control diverse physiological processes. In this Review, we summarize the recent findings of multiple groups that converged on the identification of ILCs and discuss our current understanding of the developmental and functional heterogeneity of this cell population. Furthermore, we highlight the emerging roles of ILCs in controlling tissue homeostasis in the context of infectious diseases, chronic inflammation, metabolic homeostasis and cancer.

## Definition of ILC subsets

The ILC family encompasses not only classic cytotoxic natural killer (NK) cells (which were discovered<sup>6</sup> in 1975 and are involved in protection against certain viruses and tumours) and lymphoid tissue inducer (LTi) cells (which were discovered<sup>7</sup> in 1997 and promote the formation of lymph nodes during embryogenesis), but also more recently described non-cytotoxic ILC populations. All members of the ILC family are characterized by a classic lymphoid cell morphology, but lack the expression of cell-surface molecules that identify other immune cell types, and so are defined as cell lineage marker-negative (Lin<sup>−</sup>) cells<sup>3,4,7–11</sup>. ILCs express subunits of cytokine receptors including interleukin (IL)-2 receptor- $\alpha$  (also called CD25) and IL-7 receptor- $\alpha$  (also known as CD127), but unlike adaptive T and B lymphocytes, ILCs lack expression of somatically rearranged antigen

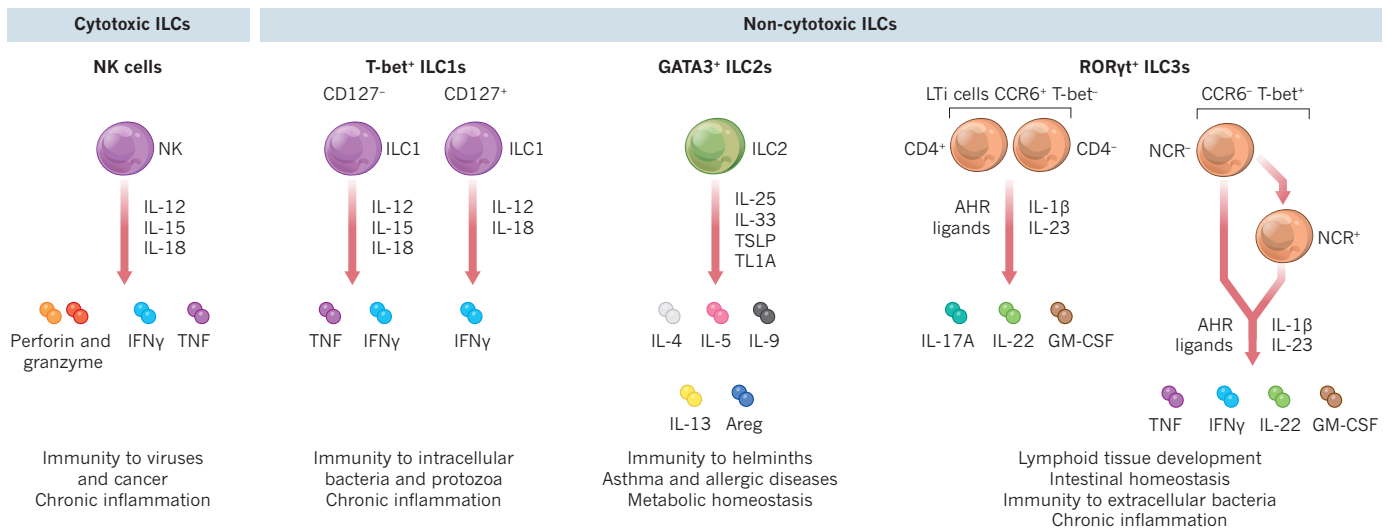
receptors and so do not exhibit any degree of antigen specificity<sup>3–5</sup>. Recent analysis of the developmental pathways of the ILC family members indicates that NK cells and non-cytotoxic helper ILCs are separate lineages. The non-cytotoxic ILCs consist of three distinct groups: group 1 ILCs (ILC1s), group 2 ILCs (ILC2s) and group 3 ILCs (ILC3s) including LTi cells<sup>3,12–14</sup> (Fig. 1). As described later, the non-cytotoxic ILC subsets are defined on the basis of their differential requirements for transcription factors during development, their patterns of expression of effector cytokines and the acquisition of other distinct effector functions<sup>3,4</sup>.

ILC1s are largely non-cytotoxic Lin<sup>−</sup> cells that are capable of producing interferon- $\gamma$  (IFN $\gamma$ ) and tumour necrosis factor (TNF) and that have been implicated in immunity to intracellular bacteria and parasites<sup>13,15–17</sup>. By contrast, ILC2s produce T helper 2 (T<sub>H</sub>2)-cell-associated cytokines (including IL-4, IL-5, IL-9 and IL-13) and/or the epidermal growth factor receptor ligand amphiregulin and promote type 2 inflammation required for anti-helminth immunity, allergic inflammation and tissue repair<sup>8–10,18–20</sup>. Finally, depending on the stimulus, ILC3s produce IL-17A, IL-17F, IL-22, granulocyte macrophage (GM) colony-stimulating factor (CSF) and TNF and can promote antibacterial immunity, chronic inflammation or tissue repair<sup>21–25</sup>. ILC3s are heterogeneous in both mice and humans. In mice, two subsets can be distinguished on the basis of their expression of the chemokine receptor CCR6. CCR6<sup>+</sup> ILC3s encompass CD4<sup>+</sup> and CD4<sup>−</sup> LTi cells. The CCR6<sup>−</sup> ILC3 population consists of two subpopulations that can be distinguished based on expression patterns of the natural cytotoxicity receptor (NCR) NKp46. In humans, almost all ILC3s express CCR6 and CD117 (also known as c-kit), and at least two subsets can be distinguished on the basis of the expression of the NCR NKp44 (refs 21–23). Strikingly, the ILC1, ILC2 and ILC3 subsets exhibit remarkable functional similarity with T-helper-cell subsets in terms of cytokine expression and potential effector function, although these cells perform their diverse functions in the absence of antigen specificity.

## Developmental requirements for ILCs

It is well established that all lymphocytes arise from a common lymphoid progenitor (CLP) that differentiates into precursors that are committed to particular cell lineages<sup>13,26,27</sup>. Whereas precursor cell populations committed to the T- and B-cell lineages have been extensively

<sup>1</sup>Weill Cornell Medical College, Cornell University, New York, New York 10021, USA. <sup>2</sup>Academic Medical Center at the University of Amsterdam, 1105 AZ Amsterdam, the Netherlands.



**Figure 1 | The innate lymphoid cell family.** Group 1, group 2 and group 3 innate lymphoid cells (ILCs) are defined by differential expression of cell-surface markers, transcription factors and patterns of expression of effector cytokines. ILCs can be activated by a diverse range of stimuli including neuropeptides, hormones, eicosanoids, cytokines and other alarmins, and can contribute to immunity, inflammation and maintenance of tissue homeostasis.

Dysregulated ILC responses can also contribute to chronic inflammatory diseases, metabolic disorders and cancer. AHR, aryl hydrocarbon receptor; Areg, amphiregulin; GM-CSF, granulocyte macrophage colony-stimulating factor; IFN $\gamma$ , interferon- $\gamma$ ; IL, interleukin; LTi, lymphoid tissue inducer; NCR, natural cytotoxicity receptor; NK, natural killer; TNF, tumour necrosis factor; TSLP, thymic stromal lymphopoietin.

characterized, CLP-derived committed ILC precursors have only recently been identified. Precursors that can develop into all ILC subsets and NK cells (but not into T cells and B cells) are contained within a population that seems to be very similar to that of CLPs but that express the integrin  $\alpha 4\beta 7$ , referred to as  $\alpha$ -lymphoid precursor ( $\alpha$ LP) cells.  $\alpha$ LP cells expressing the chemokine receptor CXCR6 are heterogeneous and can develop into conventional NK (cNK) cells and ILC3s, but not T cells or B cells and may include the common ILC and cNK cell precursors<sup>13,28,29</sup>. Downstream of these cells are at least two precursors that express the transcriptional repressor Id2, which can develop into ILCs and NK cells (see ‘Transcriptional regulation of development and function’)<sup>13,14,28</sup>. One of these Id2<sup>+</sup> ILC precursor populations expressed CD127 and the integrin  $\alpha 4\beta 7$ , and could be distinguished from CLPs by the absence of FLT3 and CD93 (ref. 13). The Id2<sup>+</sup> ILC precursors could develop into ILC1s, ILC2s and ILC3s, including LTi cells, but were unable to develop into cNK cells<sup>13</sup>.

Another ILC precursor was identified on the basis of expression of the transcription factor promyeloid leukaemia zinc finger (PLZF; encoded by *Zbtb16*), which is important for the maturation of NK T cells<sup>30,31</sup>. PLZF<sup>+</sup>CD3 $\epsilon$ <sup>-</sup> cells were found in fetal liver and adult bone marrow, expressed CD127,  $\alpha 4\beta 7$ , Thy1 and CD117, and could give rise to CD127<sup>+</sup> ILC1s, ILC2s and ILC3s *in vitro* and *in vivo* but were unable to develop into LTi cells or cNK cells<sup>14</sup>. These data suggest that Id2<sup>+</sup>PLZF<sup>+</sup> LTi cell or ILC precursors are upstream of an Id2<sup>+</sup>PLZF<sup>+</sup> ILC precursor population. Thus, our present knowledge indicates that there are at least three populations of ILC precursors with progressively limited precursor potential (Fig. 2). A bone-marrow-resident precursor that is committed to the ILC2 lineage has also been identified<sup>32,33</sup>, and there is also evidence for the existence of precursors of committed ILC1 and ILC3 subsets in mice<sup>13,28,34</sup>. Further research will be required to refine the precursors and developmental checkpoints in the formation of the three ILC subsets, and additional analyses of people with primary immunodeficiencies may provide more insight into the developmental requirements of human ILCs.

### Transcriptional regulation of development and function

The transcription factors and related molecules that control the development of ILCs are being identified at a rapid pace, building on our knowledge of transcriptional control of T-helper-cell subsets. Two transcription factors that affect early differentiation of ILCs and NK cells

have been identified, TOX and NFIL3 (refs 14, 16, 35–38). These transcription factors do not affect early development of T cells and B cells, although TOX is important at later stages of T-cell development for the differentiation of CD4<sup>+</sup> T cells<sup>39</sup>. Id2 is required for the development of ILCs at an early stage, as Id2-deficient mice lack ILCs and NK cells<sup>8,19,28,33,40</sup>. One pathway by which Id2 may allow ILC development to proceed is through sequestration of the basic helix–loop–helix transcription factor E47, which was shown to block development of LTi and NK cells<sup>41</sup>, although it has yet to be determined whether this pathway also functions in the development of ILC1s and ILC2s. It should be noted that although Id2 is required for optimal development of NK cells, this factor seems to act at a later developmental stage in the transition of pre-NK to immature NK cells<sup>13,40,41</sup>.

TCF-1 and GATA3 also seem to be key transcription factors that drive the development of all CD127<sup>+</sup> ILCs<sup>12,42–44</sup>. Genetic ablation of GATA3 in haematopoietic stem cells (HSCs) resulted in the inhibition of all ILC subsets and T cells but not of NK cells or B cells<sup>12</sup>. Deletion of GATA3 in mature ILCs affected ILC2, but not ILC3 populations<sup>12,33,45,46</sup>, suggesting that GATA3 is crucial for the post-development maintenance and survival of ILC2s. Other recent data indicate that multiple transcription factors regulate the development and function of ILC subsets. For instance, the transcription factor GFI1 might contribute to the function of GATA3 in ILC2s; this factor targets the *Gata3* gene, and levels of GATA3 expression were reduced in ILC2s deficient in GFI1 (ref. 47), providing a potential mechanism through which GATA3 expression is maintained in mature ILC2s. These findings indicate that the influence of GATA3 and associated transcriptional regulatory networks are dependent on the developmental stage of the ILC precursors, similar to their role in the development of T cells<sup>48</sup>.

The dependence of NK cells, ILC1s and ILC3s on shared transcription factors for their development and function further emphasizes the remarkable functional similarity of ILCs with corresponding T-helper-cell subsets. For example, eomesodermin and T-bet have both redundant and non-redundant roles in the development, function and migration of NK and CD8<sup>+</sup> T cells<sup>49</sup>, whereas the master transcription factor of T<sub>H</sub>1 cells, T-bet<sup>50</sup>, is important for the development and IFN $\gamma$ -producing function of non-cytotoxic ILC1s. It should also be noted that T-bet is required for upregulation of Nkp46 expression and IFN $\gamma$  production by CCR6<sup>-</sup> ILC3s<sup>13,51,52</sup>. RORyt is

indispensable for the development of all ILC3 subsets<sup>11,23,53</sup>, as it is for T<sub>H</sub>17 cells<sup>54</sup>. In addition, the aryl hydrocarbon receptor (AHR), a ligand-activated transcription factor that acts as a sensor for multiple exogenous and endogenous compounds including toxins, tryptophan metabolites, dietary products and bacterial pigments, controls the survival and function of ILC3 subsets<sup>55,56</sup> and T<sub>H</sub>17 cells<sup>57</sup>. Finally, Notch signalling, which is required for T-cell development, has also been implicated in the development of ILC3s<sup>28,29,56</sup> and ILC2s<sup>42,58</sup>. It is important to note that there are also transcription factors that act on ILC subsets but not on the corresponding T-helper cell subset. For example, ROR $\alpha$  is important for ILC2s but does not seem to play a T-cell intrinsic part in T<sub>H</sub>2 cell development and function<sup>32,58</sup>.

As with T-helper-cell subsets, there is some evidence that ILCs can show functional plasticity in response to environmental cues. In mouse models it was shown that the function of ILC3s can be affected by a gradient of expression of the transcription factors ROR $\gamma$ t and T-bet<sup>51</sup>. Following activation by cytokines such as IL-12 and IL-18, ILC3s exhibit increased expression of T-bet and decreased expression of ROR $\gamma$ t, resulting in increased IFN $\gamma$  production and loss of their capacity to produce IL-17 and IL-22 (refs 17, 51). These murine cells, termed ex-ROR $\gamma$ t<sup>+</sup> ILC3s, demonstrate ILC1-like function. In humans, an apparently similar switch in ILC3 to ILC1 effector function has been observed<sup>15,59</sup>. Whether human or mouse ILC2s exhibit plasticity in their effector functions and what the functional significance of this might be remains unclear at present.

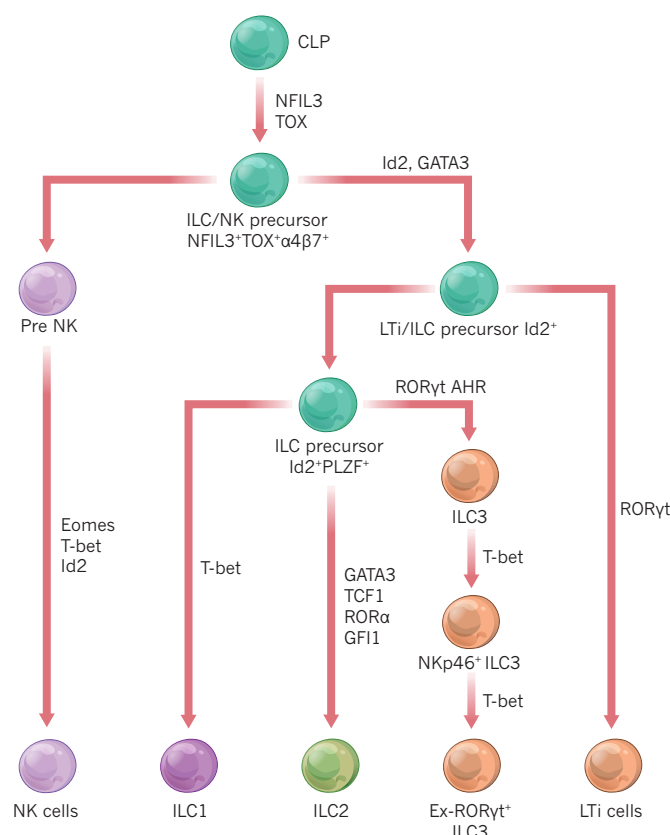
### Heterogeneous functions of ILCs

Emerging studies have identified diverse functions for ILCs, including promoting host defence against infection and regulating interactions with the microbiota. In addition, ILCs can orchestrate wound healing and tissue repair, whereas in other circumstances they can promote inflammation and tumour progression.

### ILCs promote immunity to infection

ILCs are enriched at barrier surfaces that are common sites of colonization or invasion by pathogens. It is now clear that all ILC subsets can have an important role in innate immune responses to viruses, bacteria, fungi and both intracellular and extracellular parasites at these barrier surfaces. ILC recruitment to barrier tissues takes place during embryonic development and further migration of ILCs is likely to occur in the context of ongoing inflammation. Constitutively present within adult tissues, ILCs are poised for rapid activation by host-derived cytokines and growth factors<sup>3–5</sup>. Epithelial cells and myeloid cell lineages act cooperatively to sense infection and/or tissue damage and produce cytokines and alarmins that mobilize the rapid recruitment of distinct ILC populations<sup>3</sup>. For example, IL-12, IL-15 and IL-18 activate NK cells and ILC1s<sup>13,15,16</sup>, whereas IL-2 (refs 60, 61), IL-4 (refs 62, 63), IL-25, IL-33 (refs 8–10, 19, 64–67), thymic stromal lymphopoietin (TSLP)<sup>46,68</sup>, IL-9 (refs 20, 69) and TL1A<sup>70,71</sup> activate ILC2s. By contrast, IL-1 $\beta$  and IL-23 stimulate ILC3 responses<sup>59,72</sup> (Fig. 1).

The role of NK cells in killing virus-infected cells through granzyme and/or perforin-mediated cytotoxicity is well established and has been reviewed elsewhere<sup>73</sup>. In the context of immunity to intracellular bacteria and parasites, there are important roles for ILC1s and ILC3s in host defence. For example, IFN $\gamma$ -producing ILC1s contribute to resistance to *Salmonella enterica* subsp. *enterica* serovar Typhimurium and *Toxoplasma gondii* infection in the intestine<sup>13,51</sup> (Fig. 3). In addition, before the development of adaptive immune responses, innate immunity to the extracellular Gram-negative bacterium *Citrobacter rodentium* is critically dependent on ILC3-derived IL-22 (refs 23, 74), which has an important role in promoting STAT3-dependent expression of antimicrobial peptides and maintenance of intestinal epithelial barrier function (Fig. 3)<sup>74,75</sup>. Mice deficient in IL-22 exhibited exaggerated intestinal inflammation and erosion of the epithelial barrier and rapidly succumbed to infection<sup>23,74</sup>.

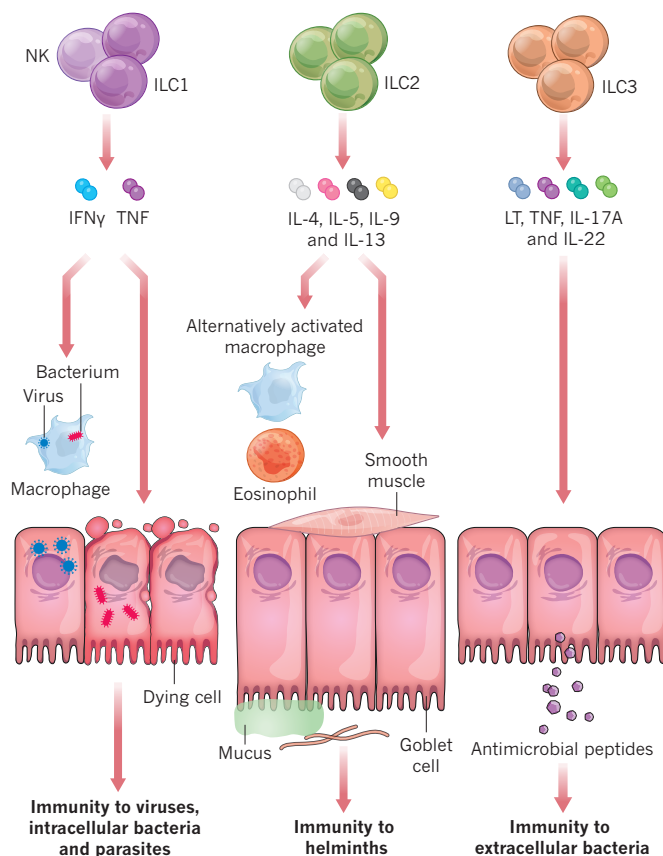


**Figure 2 | Model of developmental pathways of innate lymphoid cells and conventional natural killer cells.** Like all lymphocytes, innate lymphoid cells (ILCs) are derived from a common lymphoid progenitor (CLP). The common ILC or natural killer (NK) cell precursor is enclosed in a population of cells that have the same phenotype as the CLP but that also express the α4β7 integrin. Downstream of the common ILC/NK cell precursor (preNK) is an Id2-expressing precursor that can give rise to all ILCs, including lymphoid tissue inducer (LTi) cells and an Id2<sup>+</sup>PLZF<sup>+</sup> precursor that is restricted to ILCs but is unable to develop into LTi cells. In mice, ROR $\gamma$ t<sup>+</sup> ILC3s require T-bet to develop into interferon- $\gamma$  (IFN $\gamma$ )-producing cells. Cytokines that upregulate T-bet such as interleukin (IL)-12 and IL-18 lead to downregulation of ROR $\gamma$ t, abrogation of IL-22 and increase of IFN $\gamma$ -producing capacity. These IFN $\gamma$ -producing ILC1-like cells are also called ex-ROR $\gamma$ t<sup>+</sup> ILC3s. AHR, aryl hydrocarbon receptor.

ILC3-derived IL-22 also works cooperatively with lymphotoxin to induce fucosylation of intestinal epithelial cells, contributing to resistance to *S. Typhimurium* infection<sup>76</sup>. In addition, ILC3s play an important part in regulating host–commensal-bacteria relationships (see ‘Interactions of ILCs with the microbiota’). The host protective effects of ILC3s are not limited to bacterial infection in the intestine. ILC3s have also been implicated as an important source of IL-22 in the lungs following infection with the fungus *Candida albicans*<sup>77</sup> or the bacterium *Streptococcus pneumoniae*<sup>78</sup>, and an innate source of IFN $\gamma$  and IL-17 that resembles ILC3s in the lungs of mice subjected to a bacillus Calmette–Guérin (BCG) vaccination protocol provided enhanced protection against challenge with *Mycobacterium tuberculosis*<sup>79</sup>.

Whereas ILC1s and ILC3s promote innate immunity to viruses, intracellular bacteria and parasites, and fungi (Fig. 3), ILC2s are essential to the promotion of type 2 inflammation required for immunity to some extracellular helminth parasites<sup>1,3,4</sup>. The type 2 inflammatory response is characterized by production of cytokines, including IL-4, IL-5, IL-9 and IL-13 that regulate the alternative activation of macrophages, granulocyte responses, goblet cell hyperplasia and smooth muscle contractility that promote parasite expulsion and associated tissue-repair processes<sup>1</sup>. In studies of the mouse-adapted intestinal nematode parasite *Nippostrongylus brasiliensis*, ILC2s were





**Figure 3 | Host-protective effector functions of innate lymphoid cells at barrier surfaces.** In addition to their functions in lymphoid tissue development and metabolic homeostasis, innate lymphoid cells (ILCs) can orchestrate multiple antimicrobial effector functions at barrier surfaces in the context of exposure to viruses, bacteria, protozoa and helminths. Both natural killer (NK) cells and ILC1s produce interferon- $\gamma$  (IFN $\gamma$ ) and contribute to protective immunity against viruses, intracellular bacteria and protozoan parasites. Production of type-2 cytokines, including interleukin (IL)-4, IL-5, IL-9 and IL-13 promote alternative activation of macrophages, eosinophilia, goblet-cell hyperplasia and smooth-muscle contractility that contribute to expulsion of helminth parasites. ILC3s produce IL-17A, IL-22, lymphotoxin (LT) and tumour necrosis factor (TNF) and contribute to control of extracellular bacterial infection.

identified as the dominant non-T-cell source of IL-13, which is crucial for the expulsion of *N. brasiliensis*<sup>8–10</sup>. ILC2-deficient mice failed to efficiently expel their parasites, but adoptive transfer of wild-type ILC2s into ILC- or IL-13-deficient mice was sufficient to restore efficient worm expulsion, suggesting that these cells are crucial for the development of protective immune responses to *N. brasiliensis* in the intestine<sup>8–10</sup>.

During *N. brasiliensis* infection, IL-25 and IL-33 (refs 8–10), the signalling adaptor molecule Act1 (ref. 80) and the transcription factors GATA3 (ref. 81), TCF-1 (ref. 42) and GFI1 (ref. 47) have been shown to be essential for the population expansion and IL-13 expression by ILC2s that contribute to worm expulsion. These data suggest that a complex network of regulatory factors promote optimal ILC2 responses that directly contribute to protective immunity to helminths. In addition, recent work has shown that ILC2s interact with adaptive immune cells to indirectly promote protective type 2 immune responses. For example, ILC2s express major histocompatibility complex (MHC) class II and can activate T cells (albeit less efficiently than dendritic cells) to induce IL-2, which in turn elicits ILC2 proliferation and production of T<sub>H</sub>2-associated cytokines that promote worm expulsion<sup>61</sup>. Thus, ILC2s participate in both innate and adaptive immune responses to directly and indirectly facilitate

expulsion of *N. brasiliensis*. There is evidence that ILC2 populations also expand and contribute to protective immunity following infection with other helminth species, although the role of ILC2s in expulsion of these helminths remains less well defined. ILC2s can express amphiregulin<sup>19</sup>, which is required for immunity to the gastrointestinal nematode parasite *Trichuris muris*<sup>82</sup>, and IL-33 promotes the expansion of ILC2-like cells in the lungs of mice infected with the helminth *Strongyloides venezuelensis*<sup>83</sup>. Collectively, these studies provide the foundation for further investigation required to comprehensively define the mechanistic role of ILC2s in immune responses that coordinate protective immunity to diverse species of helminth parasites.

### Interactions of ILCs with the microbiota

In contrast to their role in promoting antimicrobial responses to pathogens, ILCs also regulate interactions between the host and the diverse array of commensal bacterial species that constitute the microbiota. ILCs do so by regulating non-haematopoietic and haematopoietic cell functions to limit inappropriate immune responses to commensal bacteria<sup>3–5</sup>. Although the accumulation of most ILC populations in the murine intestinal tissues and gut-associated lymphoid tissues can occur independently of colonization by the microbiota<sup>19,56,84–86</sup>, ILC3s seem to have a crucial role in the anatomical containment of lymphoid tissue-resident commensal bacteria<sup>86</sup>. For example, in the intestine, loss of ILC3s was associated with reduced expression of IL-22 and lower levels of antimicrobial peptides expressed by intestinal epithelial cells. These effects were coincident with the dissemination of bacteria that belong to the genus *Alcaligenes* and the development of low-grade systemic inflammation<sup>86</sup>. Intestinal ILC3s also interact with various immune cells to prevent dissemination of commensal bacteria or limit inappropriate immune responses to them. For example, ILC3s can activate B cells through membrane lymphotoxin  $\alpha 1\beta 2$  and induce the production of immunoglobulin A (IgA) that subsequently regulates dendritic cell activity and promotes immunological exclusion of commensal bacteria<sup>87</sup>. These data suggest that ILC3-dependent maintenance of intestinal epithelial and immune-cell responses limits the dissemination of select commensal bacterial species and maintains appropriate separation between inflammatory microbial-derived products and the host immune system.

ILC3s can also promote an immunologically tolerogenic state in the intestine that limits the magnitude of potentially damaging T-cell responses against commensal bacteria. ILC3s express MHC class II and are able to process and present antigens to T cells, although they lack expression of the co-stimulatory molecules required to activate T-cell responses<sup>88</sup>. In this context, depletion of ILC3s or selective deletion of MHC class II on ILC3s was associated with exaggerated commensal-bacteria-specific T<sub>H</sub>17 cell responses and, in some circumstances, the development of intestinal inflammation, indicating that intestinal ILC3s dampen commensal bacteria-specific T-cell responses in an MHC-class-II-dependent manner<sup>76,88,89</sup>. The antigen-presenting capacity of ILC3s may depend on the tissue and activation state of the cell because, in contrast to intestinal ILC3s, spleen-derived ILC3s that are cultured on feeder cells *ex vivo* can express co-stimulatory molecules and prime CD4<sup>+</sup> T cells *in vitro*<sup>90</sup>. In addition, T cells can regulate the population size and function of ILC3s in the intestine<sup>91</sup>, suggesting that dynamic cross-regulation exists between T cells and ILC3s in this tissue site. Production of GM-CSF is another mechanism by which ILC3s can promote a tolerogenic state. A recent study demonstrated that cues from the intestinal microbiota elicit GM-CSF from ILC3s to promote intestinal homeostasis through enhancing dendritic cell and regulatory T-cell function<sup>72</sup>. Human and murine ILC3s can regulate the proliferation of B cells, particularly of marginal zone B cells<sup>92</sup>, that could also contribute to immunological exclusion and anatomical containment of commensal bacteria. Collectively, the effects of ILC3s on the epithelium, dendritic cells, T cells and B cells supports the capacity of the mammalian host to tolerate colonization by the microbiota.

### ILC-mediated tissue remodelling, healing and repair

ILCs also contribute to the maintenance of tissue homeostasis by contributing to tissue remodelling, wound healing and repair processes at multiple tissues sites<sup>3,4</sup>. LT $\alpha$  cells, a subset of ILC3s, have long been recognized for their ability to promote tissue modelling in the fetus and after birth<sup>5,7,11</sup>. During embryonic development, LT $\alpha$  cells promote the formation of secondary lymphoid organs such as Peyer's patches in the gut<sup>11</sup>. LT $\alpha$  cells produce LT $\alpha$ 1 $\beta$ 2 that binds to LT $\beta$ R on stromal cells, resulting in stromal cell secretion of chemokines (CXCL13, CCL21 and CCL19) and upregulation of adhesion molecules (VCAM1, MadCam1 and ICAM1) that attract and bind leukocytes to ultimately form lymphoid structures<sup>93</sup>. After birth, LT $\alpha$  cells use the LT $\alpha$ 1 $\beta$ 2–LT $\beta$ R pathway to aid in the formation of isolated lymphoid follicles, structures that are important for immune reactions in the gut<sup>87</sup>.

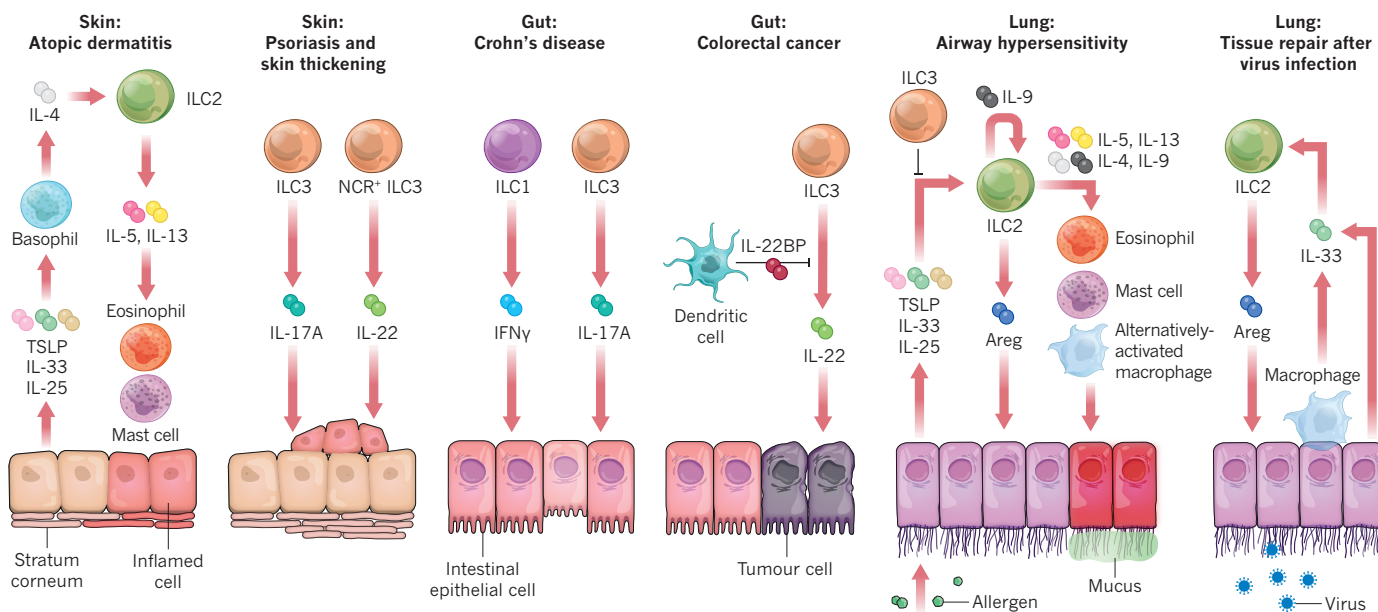
Other tissue remodelling functions of ILCs include wound healing and repair of damaged tissues. ILC3s have been implicated in repair of lymphoid tissues damaged following acute viral infection and an ensuing cytotoxic T-cell response against the virus-infected lymph node stromal cells<sup>94</sup>. IL-22-producing ILC3s also promote tissue repair and regeneration in the inflamed intestine<sup>95</sup> and in radiation-damaged thymic tissue<sup>96</sup>. Furthermore, they can limit allergic airway hyper-responsiveness in the lung<sup>97</sup>, indicating that ILC3s participate in the maintenance of homeostasis in multiple tissues following inflammation or damage. In addition, ILC2s express amphiregulin, a member of the epidermal growth factor family, and can restore bronchial epithelium that has been damaged by infection with influenza virus<sup>19</sup> (Fig. 4). Together, these studies suggest that ILC2s and ILC3s contribute to the diverse processes of tissue remodelling that promote tissue repair and homeostasis.

### ILCs and chronic inflammatory diseases

In addition to their ability to promote tissue homeostasis, ILCs can also promote inflammation at mucosal and barrier surfaces. In this context, chronic ILC activation can contribute to pathology in a wide

range of inflammatory disorders<sup>3,4</sup>. In mice, intra-epithelial ILC1s and IFN $\gamma$ -producing ILC3s can induce inflammation, and blocking IFN $\gamma$  could ameliorate this disease in some models of colitis<sup>16,25</sup>. IFN $\gamma$ -producing ILCs may also be involved in human inflammatory bowel disease because the population of ILC1s expands and IL-22-producing ILC3s decrease in inflamed intestinal tissues of patients with Crohn's disease<sup>15,16</sup>. IL-17-producing ILC3s have been shown to play a part in inflammatory bowel disease in T-cell-independent mouse models<sup>25,98</sup>. In one of these models, loss of T-bet resulted in a substantial population expansion of IL-17-producing ILCs, driven by TNF and IL-23 produced by dendritic cells<sup>98</sup>. In this context, CD127 blockade reduced intestinal ILC3 numbers and attenuated disease<sup>98</sup>. Together, these findings indicate that ILC1s and ILC3s can contribute to the development of inflammation in the intestine (Fig. 4).

ILC2s have been shown to be detrimental in a large variety of inflammatory disorders in experimental animals<sup>3,4</sup>. In mice with allergic lung inflammation, ILC2 populations expand in response to cytokines and alarmins, including IL-25, TSLP, IL-33, eicosanoids and TL1A produced by type 2 pneumocytes, alveolar macrophages and granulocytes; ILC2s also contribute to type 2 inflammation in the lung through production of IL-4, IL-5, IL-9 and IL-13 (refs 47, 70, 71, 99–101). Recent work has also suggested that ILC2s coordinate with dendritic cells and CD4<sup>+</sup> T cells at mucosal sites to shape T-cell responses that contribute to allergic airway inflammation<sup>102,103</sup>. It is unclear to what extent ILC2s are involved in human lung inflammation, but it is striking that ILC2s are increased in the peripheral blood of people with asthma relative to those without the condition<sup>104</sup>, and several genes associated with ILC2s are enriched in people with asthma, including the gene that encodes IL1RL1 (a component of the IL-33 receptor), ROR $\alpha$  and IL-13, suggesting a role for ILC2s in disease pathogenesis<sup>105</sup>. Human ILC2s are robustly expanded in another type 2 inflammatory disease, chronic rhinosinusitis, which is characterized by pronounced eosinophilia and high IgE levels in serum<sup>18,46</sup>. Epithelial cells of resected nasal polyps are capable of



**Figure 4 | Pro-inflammatory and tissue reparative functions of innate lymphoid cells.** Innate lymphoid cells (ILCs) are amplified in a variety of inflammatory diseases that affect barrier functions, suggesting that they contribute to pathology. In the skin disorder atopic dermatitis, ILC2s increase in numbers. NKp44<sup>+</sup> ILC3s (NCR<sup>+</sup>) are amplified in skin lesions of people with psoriasis and produce interleukin (IL)-22, possibly contributing to acanthosis (skin thickening), which is characteristic for this disease. IL-17-producing ILCs may increase skin inflammation. Interferon- $\gamma$  (IFN $\gamma$ )- and IL-17-producing ILCs may contribute to inflammatory bowel disease in mice, and IFN $\gamma$ -producing ILC1 are strongly amplified in inflamed intestinal tissues of

patients with Crohn's disease. ILC3-derived IL-22 promotes proliferation of tumour cells in a mouse model of infection-induced colorectal cancer. IL-22 binding protein (BP) secreted by dendritic cells can counteract the effect of IL-22. ILC2s cause airway hypersensitivity in a variety of mouse models of allergic asthma. Airway epithelial cells triggered by allergens produce thymic stromal lymphopoietin (TSLP), IL-25 and IL-33, which activate ILC2s to produce IL-5, IL-4, IL-9 and IL-13 that lead to airway hyper-reactivity. In one mouse model, ILC3s have been shown to dampen ILC2 hyper-reactivity. Lung ILC2s can mediate repair of tissue damaged by a virus through production of amphiregulin (Areg).

producing large amounts of TSLP and IL-33 that may lead to the enhanced IL-5 and IL-13 production by ILCs observed in patients with this disease<sup>18,46</sup>.

Another prominent type 2 disease is atopic dermatitis, an inflammatory skin condition characterized by the presence of  $T_H2$  cells and high levels of IL-5 and IL-13 in the skin. Recent work has suggested that ILC2s might be involved in the pathogenesis of atopic dermatitis, because the ILC2 population is expanded in the lesions of patients with atopic dermatitis<sup>67,68</sup>. These cells are stimulated by TSLP<sup>68</sup> and are also likely to be stimulated by IL-25 and IL-33, because increased transcripts of these cytokines have been detected in atopic dermatitis skin lesions<sup>65,67,68</sup>. Notably, the lectin inhibitory receptor KLRG1 may have a regulatory role in the control of ILC2 activation during atopic dermatitis, because the KLRG1 ligand epithelial cadherin (E-cadherin) diminishes ILC2 cytokine production, and E-cadherin is downregulated in atopic dermatitis lesions<sup>67</sup>. Studies in mouse models of atopic dermatitis support the idea that ILC2s can play a T-cell-independent part in establishing skin lesions<sup>68</sup>. In addition, ILC2s can interact with other innate immune cell types such as mast cells and basophils to promote type 2 inflammation in the skin<sup>62,63,106</sup> (Fig. 4).

Similar to the role for ILC2s in regulating type 2 inflammation in the skin, IL-17- and IL-22-producing ILC3s have been associated with the inflammatory skin disease psoriasis vulgaris. ILC2s and NKp44<sup>+</sup> ILC3s but not NKp44<sup>+</sup> ILC3s are present in the peripheral blood and skin of healthy people, whereas the blood and inflamed skin of those with psoriasis contained NKp44<sup>+</sup> ILC3s, indicating that dynamic changes in ILC population structure are associated with psoriatic inflammation<sup>107</sup>. In support of this idea, one patient who showed a reduction of psoriatic plaques in response to treatment with anti-TNF monoclonal antibodies also showed a decrease in circulating NKp44<sup>+</sup> ILC3s<sup>107</sup>. Studies in mouse models of psoriasis also suggest that ILC3s contribute to this disease, because mice treated with cream containing imiquimod (a Toll-like receptor 7 agonist that elicits a psoriasis-like skin inflammation) had IL-17A-, IL-17F- and IL-22-producing ILC3s and  $\gamma\delta$  T cells in the skin<sup>108</sup>.

The involvement of ILCs in various inflammatory and autoimmune disorders, as already discussed, has raised considerable interest in developing strategies to modify ILC functions to treat these diseases. Antibodies against ILC2-activating cytokines or monoclonal antibodies targeting the cytokines produced by ILC2s, including IL-5 and IL-13, are in clinical trials for the treatment of allergic diseases such as asthma and chronic rhinosinusitis<sup>109</sup>. ILC3 functions have also recently been targeted in the context of multiple sclerosis. Patients with multiple sclerosis had increased frequencies of ILC3s, and treatment with an anti-IL-2R antibody resulted in alleviation of inflammation and was associated with a reduction of ILC3s in peripheral blood<sup>110</sup>. Although this study suggests that ILC3s have a role in promoting multiple sclerosis, ILC3s seem to be dispensable in a murine model of experimental autoimmune encephalomyelitis<sup>111</sup>. In addition, small molecular compounds (SMCs) that target ILCs could also potentially be used to modify ILC functions to treat disease. In particular, arachidonic acid metabolites that regulate ILC2 function have attracted interest. Signalling by leukotriene D<sub>4</sub> to ILC2s that express the cysteinyl leukotriene receptor 1 (CysLT1R) results in secretion of type 2 cytokines, which can be inhibited by montelukast, a leukotriene receptor antagonist<sup>101</sup>. Prostaglandin D<sub>2</sub> (PGD<sub>2</sub>; which is produced by mast cells) binds to one of its receptors, CRTH2, on human ILC2s<sup>18</sup> to promote migration and cytokine production, which can be blocked by CRTH2 antagonists<sup>100,112</sup>. Notably, mast cells, which are major producers of PGD<sub>2</sub>, stably interact with ILC2s, and the resulting dialogue might induce IL-13 production<sup>106</sup>. Finally, lipoxin A<sub>4</sub>, an anti-inflammatory metabolite of arachidonic acid, binds to ILC2s and inhibits IL-13 production<sup>100</sup>. SMCs that may target ILC3 activities include antagonists of ROR $\gamma$ t, which were shown to affect biological activities of ROR $\gamma$ t-dependent  $T_H17$

cells<sup>113</sup>. Although the action of these various monoclonal antibodies or SMCs might go beyond ILCs, these data suggest that targeting ILC functions might be clinically efficacious in the context of multiple disease states.

### ILC2s and metabolic homeostasis

Emerging data highlighting an association between altered ILC responses and obesity, malnutrition and metabolic homeostasis suggest that ILCs may be crucial responders to nutrient and metabolic stress. For example, a population of IL-5- and IL-13-producing ILC2s in murine white adipose tissue<sup>114,115</sup> maintains eosinophil and alternatively activated macrophage responses that limit high-fat-diet-induced obesity and insulin resistance<sup>114</sup>. Consistent with this, IL-25-elicited ILC2 responses were associated with lower weight gain following exposure to a high fat diet<sup>115</sup>, suggesting that ILC2 responses may play an important part in regulating adipocyte development and/or function in the context of obesity. Before the discovery of ILCs another report demonstrated that delivery of IL-33 to genetically obese mice could reduce adiposity and improve insulin tolerance, whereas deletion of the IL-33 receptor was associated with exaggerated high-fat-diet-induced obesity<sup>116</sup>, supporting a potential role for IL-33 and ILC2s in metabolic homeostasis.

In addition to regulating metabolic homeostasis in white adipose tissue, ILC2s seem to be poised to rapidly respond to changes in nutrient status. Effector cytokine production by ILC2s in the small intestine was suppressed following fasting, and is regulated by the circadian clock through vasoactive intestinal peptide<sup>117</sup>. Furthermore, in the setting of malnutrition elicited by vitamin A deficiency in mice, an imbalance in ILC2 and ILC3 responses was evident, resulting in defective ILC3-dependent antibacterial immunity but enhanced ILC2-dependent anti-helminth immunity<sup>118</sup>. Together, these data provoke a model whereby alterations in nutritional status directly influence ILC-dependent maintenance of tissue homeostasis and host defence against infection. However, the mechanisms through which ILCs recognize metabolic or dietary stress and how they control metabolic homeostasis remain poorly defined.

### ILCs can promote tumour development

There is ample evidence that NK cells are involved in protection against cancer in humans and experimental animals. NK cells are particularly efficacious in killing tumour cells that have lost class I MHC antigens, a process that could be exploited for cancer therapy<sup>119</sup>. However, studies that investigate the involvement of CD127<sup>+</sup> ILCs in tumour immunity are limited. One study has shown that CCR7<sup>+</sup>CD4<sup>+</sup> ILC3s promote tumour outgrowth of melanoma cells that were engineered to express the chemokine CCL21, which was associated with the development of a suppressive tumour microenvironment<sup>120</sup>. In this context, whether ILC3s can directly tolerate T cells, as has been observed in the gut<sup>88</sup>, remains to be established.

Other evidence of a role for ILCs in tumorigenesis comes from studies investigating the pro-carcinogenic roles of cytokines involved in the activation and effector functions of ILCs, including IL-23 and IL-22. IL-23 receptor expression is increased in human colorectal cancer (CRC), and transgenic expression of IL-23 in mice induced adenomatous tumours that originated in the duodenum<sup>121</sup>. In this model, a crucial role for ILC3s in tumorigenesis in lymphocyte-deficient mice was identified, although the contribution of adaptive cells remains unclear<sup>122</sup>. In other models, a possible role for IL-22 in tumorigenesis in the gut has been described. IL-22 is important for colonic epithelial cell repair, but this activity must be tightly controlled by the soluble IL-22 receptor (IL-22 binding protein, IL-22BP) to protect against tumorigenesis, because genetic ablation of IL-22BP led to uncontrolled IL-22 production that facilitated tumour development<sup>123</sup>. In addition, in a mouse model for inflammation induced by CRC, IL-22- and IL-17-producing ROR $\gamma$ t-dependent ILC3s in the colon caused inflammation and in the presence of a carcinogen promoted



development and growth of CRC<sup>124</sup>. IL-22 might also be involved in human CRC because IL-22-producing tumour-infiltrating lymphocytes, including both T cells and non-T cells, were frequently observed in CRC, and IL-22 production in the tumour was significantly higher than in non-tumour tissue sections from the same patients<sup>124</sup>. However, the exact role of IL-22 in tumour growth and the relative importance of IL-22-producing ILC3s and T<sub>H</sub>22 cells in CRC has yet to be determined.

Together, these studies suggest that ILCs can promote tumour growth through production of tumour-promoting cytokines and by creating a suppressive tumour microenvironment. However, one study has reported an ILC3-mediated regression of a skin-residing B16 melanoma cell line that was engineered to express IL-12 (ref. 125). These ILC3s produced IFN $\gamma$  and IL-17, but the mechanism of tumour rejection induced by ILC3s in this model remains unclear<sup>125</sup>. Although it is possible that tumour-derived IL-12 induced differentiation of IL-22-producing ILC3s into protective IFN $\gamma$ -producing ex-ROR $\gamma$ <sup>+</sup> ILC3s, as has been observed in the gut<sup>15,17</sup>, further studies are required to understand the plasticity and function of ILCs within the tumour microenvironment.

### Outlook and future directions

The identification of ILCs and the subsequent recognition of their diverse functions in integrating non-haematopoietic and haematopoietic cell responses have provided new insights into how innate immune responses and tissue homeostasis are regulated in health and disease and how these innate responses affect the adaptive immune system. Although present in tissues in relatively low numbers, the selective distribution of ILCs within lymphoid and non-lymphoid tissues seems to confer a remarkable ability to regulate multiple physiological processes throughout the body. A limitation of many ILC-related studies has been the inability to genetically target select ILC populations in the presence of the adaptive immune system. The development of new genetic tools will address this challenge and could also lead to the identification of previously unrecognized innate immune cell populations that will allow greater understanding of the complexity of the immune cell network. Finally, understanding how ILC responses are dysregulated in the context of infectious, metabolic and chronic autoimmune and inflammatory conditions in humans might offer therapeutic potential in the treatment of a wide range of debilitating diseases.

**Note added in proof:** Two papers recently appeared online while the current Review was in press reporting a crucial role for ILC2s in promoting the beiging of white fat and the regulation of metabolic homeostasis (J. R. Brestoff *et al.* Group 2 innate lymphoid cells promote beiging of white adipose tissue and limit obesity. *Nature* <http://dx.doi.org/10.1038/nature14115> (2014); M. -W. Lee *et al.* Activated type 2 innate lymphoid cells regulate beige fat biogenesis. *Cell* <http://dx.doi.org/10.1016/j.cell.2014.12.011> (2014)). ■

Received 26 September; accepted 4 November 2014.

1. Pulendran, B. & Artis, D. New paradigms in type 2 immunity. *Science* **337**, 431–435 (2012).
2. Iwasaki, A. & Medzhitov, R. Regulation of adaptive immunity by the innate immune system. *Science* **327**, 291–295 (2010).
3. Spits, H. *et al.* Innate lymphoid cells – a proposal for uniform nomenclature. *Nature Rev. Immunol.* **13**, 145–149 (2013).  
**This is a key review describing the consensus nomenclature for ILC subsets developed by experts in the field.**
4. Spits, H. & Di Santo, J. P. The expanding family of innate lymphoid cells: regulators and effectors of immunity and tissue remodeling. *Nature Immunol.* **12**, 21–27 (2011).
5. Eberl, G. Development and evolution of ROR $\gamma$ <sup>+</sup> cells in a microbe's world. *Immunol. Rev.* **245**, 177–188 (2012).
6. Kiessling, R., Klein, E. & Wigzell, H. "Natural" killer cells in the mouse. I. Cytotoxic cells with specificity for mouse Moloney leukemia cells. Specificity and distribution according to genotype. *Eur. J. Immunol.* **5**, 112–117 (1975).
7. Mebius, R. E., Rennert, P. & Weissman, I. L. Developing lymph nodes collect CD4<sup>+</sup>CD3<sup>+</sup>LT $\beta$ <sup>+</sup> cells that can differentiate to APC, NK cells, and follicular cells but not T or B cells. *Immunity* **7**, 493–504 (1997).
8. Moro, K. *et al.* Innate production of T<sub>H</sub>2 cytokines by adipose tissue-associated c-Kit<sup>+</sup>Sca-1<sup>+</sup> lymphoid cells. *Nature* **463**, 540–544 (2010).
9. Neill, D. R. *et al.* Nuocytes represent a new innate effector leukocyte that mediates type-2 immunity. *Nature* **464**, 1367–1370 (2010).
10. Price, A. E. *et al.* Systemically dispersed innate IL-13-expressing cells in type 2 immunity. *Proc. Natl Acad. Sci. USA* **107**, 11489–11494 (2010).  
**This article and refs 8 and 9 provide three seminal reports identifying murine ILC2s that produce type 2 cytokines and contribute to anti-helminth immunity and type 2 inflammation.**
11. Eberl, G. *et al.* An essential function for the nuclear receptor ROR $\gamma$ t in the generation of fetal lymphoid tissue inducer cells. *Nature Immunol.* **5**, 64–73 (2004).  
**This report describes for the first time the critical dependence of LT $\beta$  cells, which help to direct the formation of secondary lymphoid structures, on the transcription factor ROR $\gamma$ t.**
12. Yagi, R. *et al.* The transcription factor GATA3 is critical for the development of all IL-7R $\alpha$ -expressing innate lymphoid cells. *Immunity* **40**, 378–388 (2014).  
**This cutting edge report describes all ILCs, not just ILC2s, as being dependent on the transcription factor GATA3 for development.**
13. Klose, C. S. *et al.* Differentiation of type 1 ILCs from a common progenitor to all helper-like innate lymphoid cell lineages. *Cell* **157**, 340–356 (2014).
14. Constantinides, M. G., McDonald, B. D., Verhoef, P. A. & Bendelac, A. A committed precursor to innate lymphoid cells. *Nature* **508**, 397–401 (2014).  
**Both this article and ref. 13 are ground-breaking studies describing committed progenitor cells of ILCs.**
15. Bernink, J. H. *et al.* Human type 1 innate lymphoid cells accumulate in inflamed mucosal tissues. *Nature Immunol.* **14**, 221–229 (2013).
16. Fuchs, A. *et al.* Intraepithelial type 1 innate lymphoid cells are a unique subset of IL-12- and IL-15-responsive IFN- $\gamma$ -producing cells. *Immunity* **38**, 769–781 (2013).  
**This article, along with ref. 15, provides evidence of non-NK cell ILC1s in humans and mice.**
17. Vonarbourg, C. *et al.* Regulated expression of nuclear receptor ROR $\gamma$ t confers distinct functional fates to NK cell receptor-expressing ROR $\gamma$ <sup>+</sup> innate lymphocytes. *Immunity* **33**, 736–751 (2010).
18. Mjösberg, J. M. *et al.* Human IL-25- and IL-33-responsive type 2 innate lymphoid cells are defined by expression of CRTH2 and CD161. *Nature Immunol.* **12**, 1055–1062 (2011).  
**This is one of the first reports of ILC2s in humans, implicating their pathological role in allergic inflammation.**
19. Monticelli, L. A. *et al.* Innate lymphoid cells promote lung-tissue homeostasis after infection with influenza virus. *Nature Immunol.* **12**, 1045–1054 (2011).  
**This report provides the first description of a tissue-protective role for ILC2s, describing how ILC2s produce amphiregulin, a ligand of EGFR, and contribute to lung-tissue repair following influenza A virus infection in mice.**
20. Wilhelm, C. *et al.* An IL-9 fate reporter demonstrates the induction of an innate IL-9 response in lung inflammation. *Nature Immunol.* **12**, 1071–1077 (2011).
21. Cella, M. *et al.* A human natural killer cell subset provides an innate source of IL-22 for mucosal immunity. *Nature* **457**, 722–725 (2009).  
**This is one of the first descriptions of IL-22-producing ILC3s in humans and mice, implicating their role in anti-pathogen immunity.**
22. Cupedo, T. *et al.* Human fetal lymphoid tissue-inducer cells are interleukin 17-producing precursors to ROR $\gamma$ <sup>+</sup>CD127<sup>+</sup> natural killer-like cells. *Nature Immunol.* **10**, 66–74 (2009).
23. Satoh-Takayama, N. *et al.* Microbial flora drives interleukin 22 production in intestinal NKp46<sup>+</sup> cells that provide innate mucosal immune defense. *Immunity* **29**, 958–970 (2008).
24. Sonnenberg, G. F., Monticelli, L. A., Elloso, M. M., Fouser, L. A. & Artis, D. CD4<sup>+</sup> lymphoid tissue-inducer cells promote innate immunity in the gut. *Immunity* **34**, 122–134 (2011).
25. Buonocore, S. *et al.* Innate lymphoid cells drive interleukin-23-dependent innate intestinal pathology. *Nature* **464**, 1371–1375 (2010).  
**This paper was the first report of a role for ILC3-like cells in promoting intestinal inflammation.**
26. Ichii, M. *et al.* Functional diversity of stem and progenitor cells with B-lymphopoietic potential. *Immunol. Rev.* **237**, 10–21 (2010).
27. Yang, Q., Jeremiah Bell, J. & Bhandoola, A. T-cell lineage determination. *Immunol. Rev.* **238**, 12–22 (2010).
28. Cherrier, M., Sawa, S. & Eberl, G. Notch, Id2, and ROR $\gamma$ t sequentially orchestrate the fetal development of lymphoid tissue inducer cells. *J. Exp. Med.* **209**, 729–740 (2012).
29. Possot, C. *et al.* Notch signaling is necessary for adult, but not fetal, development of ROR $\gamma$ <sup>+</sup> innate lymphoid cells. *Nature Immunol.* **12**, 949–958 (2011).
30. Kovalovsky, D. *et al.* The BTB-zinc finger transcriptional regulator PLZF controls the development of invariant natural killer T cell effector functions. *Nature Immunol.* **9**, 1055–1064 (2008).
31. Savage, A. K. *et al.* The transcription factor PLZF directs the effector program of the NKT cell lineage. *Immunity* **29**, 391–403 (2008).
32. Halim, T. Y. *et al.* Retinoic-acid-receptor-related orphan nuclear receptor alpha is required for natural helper cell development and allergic inflammation. *Immunity* **37**, 463–474 (2012).
33. Hoyer, T. *et al.* The transcription factor GATA-3 controls cell fate and maintenance of type 2 innate lymphoid cells. *Immunity* **37**, 634–648 (2012).
34. van de Pavert, S. A. *et al.* Maternal retinoids control type 3 innate lymphoid cells and set the offspring immunity. *Nature* **508**, 123–127 (2014).

35. Aliahmad, P., de la Torre, B. & Kaye, J. Shared dependence on the DNA-binding factor TOX for the development of lymphoid tissue-inducer cell and NK cell lineages. *Nature Immunol.* **11**, 945–952 (2010).
36. Geiger, T. L. *et al.* Nfil3 is crucial for development of innate lymphoid cells and host protection against intestinal pathogens. *J. Exp. Med.* **211**, 1723–1731 (2014).
37. Seillet, C. *et al.* Nfil3 is required for the development of all innate lymphoid cell subsets. *J. Exp. Med.* **211**, 1733–1740 (2014).
38. Yu, X. *et al.* The basic leucine zipper transcription factor NFIL3 directs the development of a common innate lymphoid cell precursor. *eLife* **3**, e04406 (2014).
39. Aliahmad, P. & Kaye, J. Development of all CD4 T lineages requires nuclear factor TOX. *J. Exp. Med.* **205**, 245–256 (2008).
40. Yokota, Y. *et al.* Development of peripheral lymphoid organs and natural killer cells depends on the helix-loop-helix inhibitor Id2. *Nature* **397**, 702–706 (1999).
41. Boos, M. D., Yokota, Y., Eberl, G. & Kee, B. L. Mature natural killer cell and lymphoid tissue-inducing cell development requires Id2-mediated suppression of E protein activity. *J. Exp. Med.* **204**, 1119–1130 (2007).
42. Yang, Q. *et al.* T cell factor 1 is required for group 2 innate lymphoid cell generation. *Immunity* **38**, 694–704 (2013).
43. Serafini, N. *et al.* Gata3 drives development of RORγt<sup>+</sup> group 3 innate lymphoid cells. *J. Exp. Med.* **211**, 199–208 (2014).
44. Malhotra, N. *et al.* A network of high-mobility group box transcription factors programs innate interleukin-17 production. *Immunity* **38**, 681–693 (2013).
45. Klein Wolterink, R. G. *et al.* Essential, dose-dependent role for the transcription factor Gata3 in the development of IL-5<sup>+</sup> and IL-13<sup>+</sup> type 2 innate lymphoid cells. *Proc. Natl Acad. Sci. USA* **110**, 10240–10245 (2013).
46. Mjösberg, J. *et al.* The transcription factor GATA3 is essential for the function of human type 2 innate lymphoid cells. *Immunity* **37**, 649–659 (2012).
47. Spooner, C. J. *et al.* Specification of type 2 innate lymphocytes by the transcriptional determinant Gfi1. *Nature Immunol.* **14**, 1229–1236 (2013).
48. Ting, C. N., Olson, M. C., Barton, K. P. & Leiden, J. M. Transcription factor GATA-3 is required for development of the T-cell lineage. *Nature* **384**, 474–478 (1996).
49. Gordon, S. M. *et al.* The transcription factors T-bet and Eomes control key checkpoints of natural killer cell maturation. *Immunity* **36**, 55–67 (2012).
50. Szabo, S. J. *et al.* A novel transcription factor, T-bet, directs Th1 lineage commitment. *Cell* **100**, 655–669 (2000).
51. Klose, C. S. *et al.* A T-bet gradient controls the fate and function of CCR6-RORγt<sup>+</sup> innate lymphoid cells. *Nature* **494**, 261–265 (2013).
52. Rankin, L. C. *et al.* The transcription factor T-bet is essential for the development of NKp46<sup>+</sup> innate lymphocytes via the Notch pathway. *Nature Immunol.* **14**, 389–395 (2013).
53. Sun, Z. *et al.* Requirement for RORγ in thymocyte survival and lymphoid organ development. *Science* **288**, 2369–2373 (2000).
54. Ivanov, I. I. *et al.* The orphan nuclear receptor RORγt directs the differentiation program of proinflammatory IL-17<sup>+</sup> T helper cells. *Cell* **126**, 1121–1133 (2006).
55. Kiss, E. A. *et al.* Natural aryl hydrocarbon receptor ligands control organogenesis of intestinal lymphoid follicles. *Science* **334**, 1561–1565 (2011).
56. Lee, J. S. *et al.* AHR drives the development of gut ILC22 cells and postnatal lymphoid tissues via pathways dependent on and independent of Notch. *Nature Immunol.* **13**, 144–151 (2012).
57. Veldhoen, M. *et al.* The aryl hydrocarbon receptor links T<sub>H</sub>17-cell-mediated autoimmunity to environmental toxins. *Nature* **453**, 106–109 (2008).
58. Wong, S. H. *et al.* Transcription factor RORα is critical for nuocyte development. *Nature Immunol.* **13**, 229–236 (2012).
59. Cella, M., Otero, K. & Colonna, M. Expansion of human NK-22 cells with IL-7, IL-2, and IL-1β reveals intrinsic functional plasticity. *Proc. Natl Acad. Sci. USA* **107**, 10961–10966 (2010).
60. Mirchandani, A. S. *et al.* Type 2 innate lymphoid cells drive CD4<sup>+</sup> Th2 cell responses. *J. Immunol.* **192**, 2442–2448 (2014).
61. Oliphant, C. J. *et al.* MHCII-mediated dialog between group 2 innate lymphoid cells and CD4<sup>+</sup> T cells potentiates type 2 immunity and promotes parasitic helminth expulsion. *Immunity* **41**, 283–295 (2014).
- This paper is a cutting-edge description of how ILC2s and T cells interact to promote type 2 inflammation that drives helminth expulsion from the gut.**
62. Kim, B. S. *et al.* Basophils promote innate lymphoid cell responses in inflamed skin. *J. Immunol.* **193**, 3717–3725 (2014).
63. Motomura, Y. *et al.* Basophil-derived interleukin-4 controls the function of natural helper cells, a member of ILC2s, in lung inflammation. *Immunity* **40**, 758–771 (2014).
64. McHedlidze, T. *et al.* Interleukin-33-dependent innate lymphoid cells mediate hepatic fibrosis. *Immunity* **39**, 357–371 (2013).
65. Imai, Y. *et al.* Skin-specific expression of IL-33 activates group 2 innate lymphoid cells and elicits atopic dermatitis-like inflammation in mice. *Proc. Natl Acad. Sci. USA* **110**, 13921–13926 (2013).
66. Saenz, S. A. *et al.* IL-25 simultaneously elicits distinct populations of innate lymphoid cells and multipotent progenitor type 2 (MP2) cells. *J. Exp. Med.* **210**, 1823–1837 (2013).
67. Salimi, M. *et al.* A role for IL-25 and IL-33-driven type-2 innate lymphoid cells in atopic dermatitis. *J. Exp. Med.* **210**, 2939–2950 (2013).
68. Kim, B. S. *et al.* TSLP elicits IL-33-independent innate lymphoid cell responses to promote skin inflammation. *Sci. Transl. Med.* **5**, 170ra116 (2013).
69. Turner, J. E. *et al.* IL-9-mediated survival of type 2 innate lymphoid cells promotes damage control in helminth-induced lung inflammation.
- J. Exp. Med.* **210**, 2951–2965 (2013).
70. Meylan, F. *et al.* The TNF-family cytokine TL1A promotes allergic immunopathology through group 2 innate lymphoid cells. *Mucosal Immunol.* **7**, 958–968 (2014).
71. Yu, X. *et al.* TNF superfamily member TL1A elicits type 2 innate lymphoid cells at mucosal barriers. *Mucosal Immunol.* **7**, 730–740 (2014).
72. Mortha, A. *et al.* Microbiota-dependent crosstalk between macrophages and ILC3 promotes intestinal homeostasis. *Science* **343**, 1249288 (2014).
73. Biron, C. A., Nguyen, K. B., Pien, G. C., Cousens, L. P. & Salazar-Mather, T. P. Natural killer cells in antiviral defense: function and regulation by innate cytokines. *Annu. Rev. Immunol.* **17**, 189–220 (1999).
74. Sonnenberg, G. F., Fouser, L. A. & Artis, D. Border patrol: regulation of immunity, inflammation and tissue homeostasis at barrier surfaces by IL-22. *Nature Immunol.* **12**, 383–390 (2011).
75. Zheng, Y. *et al.* Interleukin-22 mediates early host defense against attaching and effacing bacterial pathogens. *Nature Med.* **14**, 282–289 (2008).
76. Goto, Y. *et al.* Innate lymphoid cells regulate intestinal epithelial cell glycosylation. *Science* **345**, 1254009 (2014).
77. Gladiator, A., Wangler, N., Trautwein-Weidner, K. & LeibundGut-Landmann, S. Cutting edge: IL-17-secreting innate lymphoid cells are essential for host defense against fungal infection. *J. Immunol.* **190**, 521–525 (2013).
78. Van Maele, L. *et al.* Activation of type 3 innate lymphoid cells and interleukin 22 secretion in the lungs during *Streptococcus pneumoniae* infection. *J. Infect. Dis.* **210**, 493–503 (2014).
79. Pitt, J. M. *et al.* Blockade of IL-10 signaling during bacillus Calmette-Guerin vaccination enhances and sustains Th1, Th17, and innate lymphoid IFN-γ and IL-17 responses and increases protection to *Mycobacterium tuberculosis* infection. *J. Immunol.* **189**, 4079–4087 (2012).
80. Kang, Z. *et al.* Epithelial cell-specific Act1 adaptor mediates interleukin-25-dependent helminth expulsion through expansion of Lin<sup>+</sup> c-Kit<sup>+</sup> innate cell population. *Immunity* **36**, 821–833 (2012).
81. Liang, H. E. *et al.* Divergent expression patterns of IL-4 and IL-13 define unique functions in allergic immunity. *Nature Immunol.* **13**, 58–66 (2012).
82. Zaiss, D. M. *et al.* Amphiregulin, a T<sub>H</sub>2 cytokine enhancing resistance to nematodes. *Science* **314**, 1746 (2006).
83. Yasuda, K. *et al.* Contribution of IL-33-activated type II innate lymphoid cells to pulmonary eosinophilia in intestinal nematode-infected mice. *Proc. Natl Acad. Sci. USA* **109**, 3451–3456 (2012).
84. Ganal, S. C. *et al.* Priming of natural killer cells by nonmucosal mononuclear phagocytes requires instructive signals from commensal microbiota. *Immunity* **37**, 171–186 (2012).
85. Sawa, S. *et al.* Lineage relationship analysis of RORγt<sup>+</sup> innate lymphoid cells. *Science* **330**, 665–669 (2010).
86. Sonnenberg, G. F. *et al.* Innate lymphoid cells promote anatomical containment of lymphoid-resident commensal bacteria. *Science* **336**, 1321–1325 (2012).
- This report identifies a crucial role for ILC3s in regulating host-microbiota interactions.**
87. Kruglov, A. A. *et al.* Nonredundant function of soluble LTα3 produced by innate lymphoid cells in intestinal homeostasis. *Science* **342**, 1243–1246 (2013).
88. Hepworth, M. R. *et al.* Innate lymphoid cells regulate CD4<sup>+</sup> T-cell responses to intestinal commensal bacteria. *Nature* **498**, 113–117 (2013).
- This paper provides the first evidence that ILC3s directly regulate adaptive immune responses.**
89. Qiu, J. *et al.* The aryl hydrocarbon receptor regulates gut immunity through modulation of innate lymphoid cells. *Immunity* **36**, 92–104 (2012).
90. von Burg, N. *et al.* Activated group 3 innate lymphoid cells promote T-cell-mediated immune responses. *Proc. Natl Acad. Sci. USA* **111**, 12835–12840 (2014).
91. Korn, L. L. *et al.* Conventional CD4<sup>+</sup> T cells regulate IL-22-producing intestinal innate lymphoid cells. *Mucosal Immunol.* **7**, 1045–1057 (2014).
92. Magri, G. *et al.* Innate lymphoid cells integrate stromal and immunological signals to enhance antibody production by splenic marginal zone B cells. *Nature Immunol.* **15**, 354–364 (2014).
93. van de Pavert, S. A. *et al.* Chemokine CXCL13 is essential for lymph node initiation and is induced by retinoic acid and neuronal stimulation. *Nature Immunol.* **10**, 1193–1199 (2009).
94. Scandella, E. *et al.* Restoration of lymphoid organ integrity through the interaction of lymphoid tissue-inducer cells with stroma of the T cell zone. *Nature Immunol.* **9**, 667–675 (2008).
95. Sawa, S. *et al.* RORγt<sup>+</sup> innate lymphoid cells regulate intestinal homeostasis by integrating negative signals from the symbiotic microbiota. *Nature Immunol.* **12**, 320–326 (2011).
96. Dudakov, J. A. *et al.* Interleukin-22 drives endogenous thymic regeneration in mice. *Science* **336**, 91–95 (2012).
97. Taube, C. *et al.* IL-22 is produced by innate lymphoid cells and limits inflammation in allergic airway disease. *PLoS ONE* **6**, e21799 (2011).
98. Powell, N. *et al.* The transcription factor T-bet regulates intestinal inflammation mediated by interleukin-7 receptor<sup>+</sup> innate lymphoid cells. *Immunity* **37**, 674–684 (2012).
99. Chang, Y. J. *et al.* Innate lymphoid cells mediate influenza-induced airway hyper-reactivity independently of adaptive immunity. *Nature Immunol.* **12**, 631–638 (2011).
100. Barnig, C. *et al.* Lipoxin A4 regulates natural killer cell and type 2 innate lymphoid cell activation in asthma. *Sci. Transl. Med.* **5**, 174ra126 (2013).
- This is one of the only reports to describe a factor or pathway that limits ILC2 function.**

101. Doherty, T. A. *et al.* Lung type 2 innate lymphoid cells express cysteinyl leukotriene receptor 1, which regulates  $T_H2$  cytokine production. *J. Allergy Clin. Immunol.* **132**, 205–213 (2013).
102. Drake, L. Y., Iijima, K. & Kita, H. Group 2 innate lymphoid cells and CD4<sup>+</sup> T cells cooperate to mediate type 2 immune response in mice. *Allergy* **69**, 1300–1307 (2014).
103. Halim, T. Y. *et al.* Group 2 innate lymphoid cells are critical for the initiation of adaptive T helper 2 cell-mediated allergic lung inflammation. *Immunity* **40**, 425–435 (2014).
104. Bartemes, K. R., Kephart, G. M., Fox, S. J. & Kita, H. Enhanced innate type 2 immune response in peripheral blood from patients with asthma. *J. Allergy Clin. Immunol.* **134**, 671–678 (2014).
105. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363**, 1211–1221 (2010).
106. Roediger, B. *et al.* Cutaneous immunosurveillance and regulation of inflammation by group 2 innate lymphoid cells. *Nature Immunol.* **14**, 564–573 (2013).
107. Villanova, F. *et al.* Characterization of innate lymphoid cells in human skin and blood demonstrates increase of NKp44<sup>+</sup> ILC3 in psoriasis. *J. Invest. Dermatol.* **134**, 984–991 (2014).
108. Pantelyushin, S. *et al.* Rorγt<sup>+</sup> innate lymphocytes and γδ T cells initiate psoriasiform plaque formation in mice. *J. Clin. Invest.* **122**, 2252–2256 (2012).
109. Hambly, N. & Nair, P. Monoclonal antibodies for the treatment of refractory asthma. *Curr. Opin. Pulm. Med.* **20**, 87–94 (2014).
110. Perry, J. S. *et al.* Inhibition of LT $\alpha$  cell development by CD25 blockade is associated with decreased intrathecal inflammation in multiple sclerosis. *Sci. Transl. Med.* **4**, 145ra106 (2012).  
**This paper identifies that targeting pathways that regulate ILCs, such as IL-2 signalling, can be efficacious in treating diseases associated with aberrant ILC responses.**
111. Mair, F. & Becher, B. Thy1<sup>+</sup> Sca1<sup>+</sup> innate lymphoid cells infiltrate the CNS during autoimmune inflammation, but do not contribute to disease development. *Eur. J. Immunol.* **44**, 37–45 (2014).
112. Xue, L. *et al.* Prostaglandin D<sub>2</sub> activates group 2 innate lymphoid cells through chemoattractant receptor-homologous molecule expressed on  $T_H2$  cells. *J. Allergy Clin. Immunol.* **133**, 1184–1194 (2014).
113. Solt, L. A. *et al.* Suppression of  $T_H17$  differentiation and autoimmunity by a synthetic ROR ligand. *Nature* **472**, 491–494 (2011).
114. Molofsky, A. B. *et al.* Innate lymphoid type 2 cells sustain visceral adipose tissue eosinophils and alternatively activated macrophages. *J. Exp. Med.* **210**, 535–549 (2013).
115. Hams, E., Locksley, R. M., McKenzie, A. N. & Fallon, P. G. Cutting edge: IL-25 elicits innate lymphoid type 2 and type II NKT cells that regulate obesity in mice. *J. Immunol.* **191**, 5349–5353 (2013).
116. Miller, A. M. *et al.* Interleukin-33 induces protective effects in adipose tissue inflammation during obesity in mice. *Circ. Res.* **107**, 650–658 (2010).
117. Nussbaum, J. C. *et al.* Type 2 innate lymphoid cells control eosinophil homeostasis. *Nature* **502**, 245–248 (2013).  
**This report establishes a connection between eosinophil homeostasis and ILC2 activity in the context of the regulation of food intake and metabolic homeostasis that is controlled by circadian rhythms in the intestine.**
118. Spencer, S. P. *et al.* Adaptation of innate lymphoid cells to a micronutrient deficiency promotes type 2 barrier immunity. *Science* **343**, 432–437 (2014).
119. Vivier, E., Ugolini, S., Blaise, D., Chabannon, C. & Brossay, L. Targeting natural killer cells and natural killer T cells in cancer. *Nature Rev. Immunol.* **12**, 239–252 (2012).
120. Shields, J. D., Kourtis, I. C., Tomei, A. A., Roberts, J. M. & Swartz, M. A. Induction of lymphoid-like stroma and immune escape by tumors that express the chemokine CCL21. *Science* **328**, 749–752 (2010).
121. Langowski, J. L. *et al.* IL-23 promotes tumour incidence and growth. *Nature* **442**, 461–465 (2006).
122. Chan, I. H. *et al.* Interleukin-23 is sufficient to induce rapid *de novo* gut tumorigenesis, independent of carcinogens, through activation of innate lymphoid cells. *Mucosal Immunol.* **7**, 842–856 (2014).
123. Huber, S. *et al.* IL-22BP is regulated by the inflammasome and modulates tumorigenesis in the intestine. *Nature* **491**, 259–263 (2012).
124. Kirchberger, S. *et al.* Innate lymphoid cells sustain colon cancer through production of interleukin-22 in a mouse model. *J. Exp. Med.* **210**, 917–931 (2013).
125. Eisenring, M., vom Berg, J., Kristiansen, G., Saller, E. & Becher, B. IL-12 initiates tumor rejection via lymphoid tissue-inducer cells bearing the natural cytotoxicity receptor NKp46. *Nature Immunol.* **11**, 1030–1038 (2010).

**Acknowledgements** We thank L. Osborne, K. Germar, M. R. Hepworth, E. Tait Wojno and G. F. Sonnenberg for discussions and critical reading of the manuscript. We apologize to colleagues whose work could not be directly quoted due to space constraints. Research in the Artis laboratory is supported by the US National Institutes of Health (AI061570, AI095608, AI074878, AI095466, AI106697, AI102942, AI097333), the Crohns and Colitis Foundation of America and the Burroughs Wellcome Fund. Research in the Spits lab is supported by an advanced grant (341038) of the European Research Council.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: see [go.nature.com/9rjgkc](http://go.nature.com/9rjgkc) for details. Readers are welcome to comment on the online version of this paper at [go.nature.com/9rjgkc](http://go.nature.com/9rjgkc). Correspondence should be addressed to D.A. ([dartis@med.cornell.edu](mailto:dartis@med.cornell.edu)) or H.S. ([hergen.spits@amc.uva.nl](mailto:hergen.spits@amc.uva.nl)).



# Nutrient-sensing mechanisms and pathways

Alejo Efeyan<sup>1,2,3,4</sup>, William C. Comb<sup>1,2,3,4</sup> & David M. Sabatini<sup>1,2,3,4,5</sup>

**The ability to sense and respond to fluctuations in environmental nutrient levels is a requisite for life. Nutrient scarcity is a selective pressure that has shaped the evolution of most cellular processes. Different pathways that detect intracellular and extracellular levels of sugars, amino acids, lipids and surrogate metabolites are integrated and coordinated at the organismal level through hormonal signals. During food abundance, nutrient-sensing pathways engage anabolism and storage, whereas scarcity triggers homeostatic mechanisms, such as the mobilization of internal stores through autophagy. Nutrient-sensing pathways are commonly deregulated in human metabolic diseases.**

Nutrients (also referred to as macronutrients) are simple organic compounds that are involved in biochemical reactions that produce energy or are constituents of cellular biomass. Glucose and related sugars, amino acids and lipids are important cellular nutrients, and distinct mechanisms to sense their abundances operate in mammalian cells. Essentiality is not necessarily a hallmark of nutrients; for certain amino acids, such as arginine, cysteine, glutamine, glycine, proline and tyrosine, essentiality is context dependent. In healthy people, the *de novo* synthesis of these amino acids from other molecules meets organismal requirements, but under particular metabolic needs, such as during the rapid growth of infants<sup>1,2</sup>, they must also be obtained from the environment. Nutrient scarcity has operated as a strong pressure for selecting efficient mechanisms of nutrient sensing in all species. Considering the importance of nutrient homeostasis for all living organisms, and for human health in particular, it is surprising that we know relatively little about direct nutrient-sensing mechanisms.

The sensing of a particular nutrient may involve the direct binding of the molecule to its sensor, or occur by an indirect mechanism relying on the detection of a surrogate molecule that reflects nutrient abundance. Regardless of the manner in which nutrient sensing occurs, for a protein to be considered a sensor, its affinity must be within the range of physiological fluctuations of the concentration of the nutrient or its surrogate.

Unicellular organisms are directly exposed to environmental fluctuations of nutrients, and sense both intracellular and environmental nutrient levels. By contrast, most cells in multicellular eukaryotes are not directly exposed to changes in environmental nutrients, and homeostatic responses aimed at maintaining circulating nutrient levels within a narrow range exist. Nevertheless, internal nutrient levels do fluctuate, and hence intracellular and extracellular nutrient-sensing mechanisms are also present in mammals. In multicellular organisms, nutrients also trigger the release of hormones, which act as long-range signals with non-cell-autonomous effects, to facilitate the coordination of coherent responses in the organism as a whole.

In this Review, we discuss intracellular and extracellular glucose-, amino-acid- and lipid-sensing mechanisms and signalling events in mammals; discuss how these sensing mechanisms become deregulated in human disease; and describe how internal nutrient stores are mobilized during nutrient scarcity.

## Lipid sensing

Lipids are a large and diverse set of nutrients (for example, fatty acids or cholesterol) characterized by hydrophobic carbon backbones that are used for energy storage and membrane biosynthesis, among other cellular processes. Owing to their non-polar nature, lipids are normally either packaged into lipoproteins and chylomicrons or bound by albumin in the serum<sup>3</sup>; they are rarely found free in a soluble form in the organism. Despite the morbidity caused by high levels of lipid intake and deregulated lipid storage, which occurs in obese states, our knowledge of lipid-sensing mechanisms, with some exceptions, is quite limited.

## Fatty-acid signalling

A family of G-protein-coupled receptors, best characterized by GPR40 and GPR120, detects long-chain unsaturated fatty acids. In mechanisms that are not fully understood, free fatty-acid stimulation of GPR40 at the plasma membrane of pancreatic  $\beta$ -cells augments glucose-stimulated insulin release<sup>4</sup> (Fig. 1a). GPR120 also mediates insulinotropic activity, albeit by an indirect mechanism, involving production of GLP1 in the gut and the release into circulation. GLP1 belongs to a group of gastrointestinal hormones called incretins that promote insulin release in  $\beta$ -cells<sup>5</sup>. These examples demonstrate how an increase in one particular nutrient (fatty acids) anticipates a response to the imminent increase in another nutrient (glucose), as food intake rarely provides solely one nutrient species. In addition, activation of GPR120 at the plasma membrane of white adipocytes leads to a signal transduction cascade that promotes phosphatidylinositol-3-OH kinase (PI(3)K) and AKT activation, leading to the cell-autonomous induction of glucose uptake<sup>6</sup> (Fig. 1a). Genetic mutations that disrupt GPR120 function occur in people who are obese, and ablation of *Gpr120* in mice contributes to diet-induced obesity, suggesting that this signal transduction pathway has a key role in the systemic control of nutrient homeostasis<sup>7</sup>. Naturally, these findings have spurred interest in the development of GPR120 agonists to control the onset of obesity<sup>8</sup>.

In addition to GPR120, the CD36 (also known as FAT) receptor has been implicated in direct binding and uptake of intestinal lumen fatty acids<sup>9</sup>, and, interestingly, GPR40, GPR120 and CD36 have fatty-acid-sensing properties in cells within the oral epithelium that are involved in gustatory perception<sup>10–13</sup> (Fig. 1a).

<sup>1</sup>Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA. <sup>2</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. <sup>3</sup>David H. Koch Institute for Integrative Cancer Research at Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. <sup>4</sup>Broad Institute, Seven Cambridge Center, Cambridge, Massachusetts 02142, USA. <sup>5</sup>Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

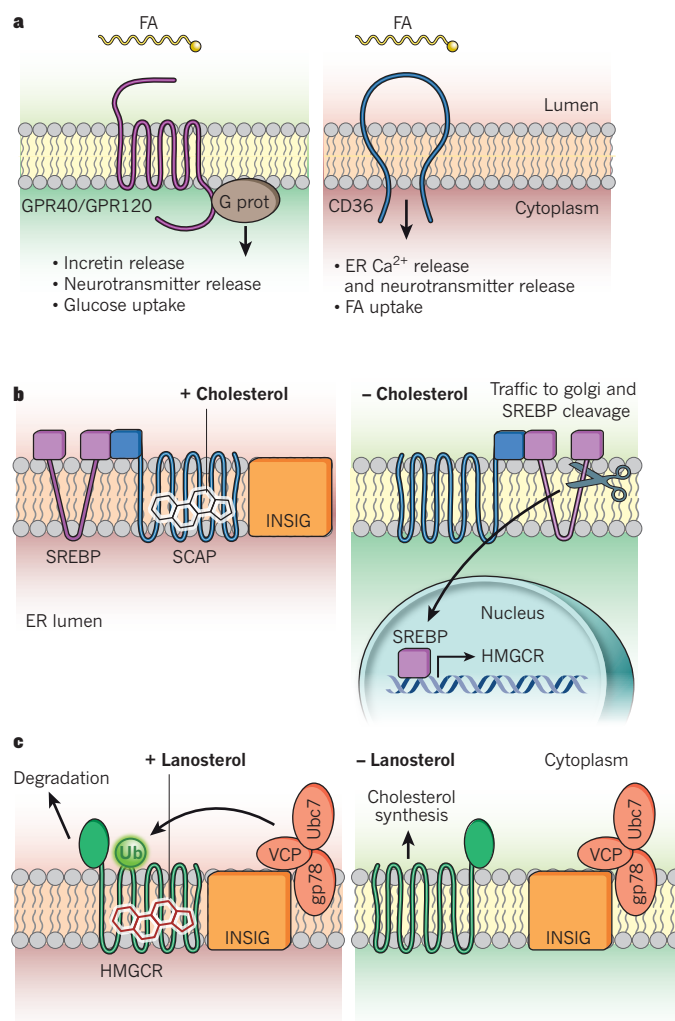
## Cholesterol sensing

Our limited knowledge about the sensing of other lipid species is in contrast with our profound understanding of the cholesterol-sensing mechanism, which was deciphered by Brown and Goldstein<sup>14</sup>. Sterols, including cholesterol, are fundamental constituents of mammalian membranes that provide membrane fluidity and are needed for the synthesis of steroid hormones. Cholesterol can be obtained from the diet, as well as synthesized *de novo*. Hence, adequate sensing of internal cholesterol levels allows the energetically demanding cholesterol biosynthetic pathway to be controlled, so that it is only active when external supply and internal levels of sterols are low. Cholesterol sensing occurs in close proximity to the regulation of the cholesterol biosynthetic pathway: the cholesterol-sensing protein (cholesterol-sensing protein SREBP1 cleavage activating protein, SCAP), and the transcription factor that induces the expression of enzymes involved in the cholesterol biosynthetic pathway, form a constitutively bound complex on the endoplasmic reticulum (ER). SCAP directly binds cholesterol by a region originally found to span its five transmembrane sterol sensing domains (SSDs)<sup>15,16</sup>. The initial mapping observations were later refined to a loop in the ER lumen side of the membrane, probably embedded in the lipid bilayer<sup>17</sup> (Fig. 1b). SCAP is constitutively bound to sterol regulatory element-binding proteins (SREBPs), which transactivate genes that are crucial for cholesterol synthesis. When cholesterol levels are high, cholesterol binding to SCAP triggers a conformational change that increases its affinity for the INSIG proteins<sup>18</sup>, an anchor for SCAP and SREBP within ER membranes. Conversely, when cholesterol levels are low and SCAP is not bound to cholesterol, the SCAP–SREBP tandem dissociates from INSIG and shuttles to the Golgi apparatus<sup>19</sup> (Fig. 1b). This step is essential because the presence of the SCAP–SREBP complex at the Golgi allows the cleavage and release of the cytoplasmic amino-terminus of SREBP by proteases that are resident at the Golgi<sup>20,21</sup>. In turn, the cleaved cytoplasmic fragment of SREBP translocates to the nucleus and induces genes involved in lipid anabolism. Replete cholesterol levels then initiate a slow negative feedback by interacting with SCAP and inhibiting further cleavage of SREBP<sup>22</sup>.

Substantial evidence supports an additional sterol-sensing event that occurs within the ER, involving the enzyme HMG-CoA reductase. HMG-CoA reductase, a transcriptional target of SREBP, catalyses the rate-limiting step in *de novo* cholesterol synthesis in response to low cholesterol levels. The carboxy-terminus of HMG-CoA reductase, containing its catalytic activity, is exposed to the cytoplasm, whereas several transmembrane domains, including the sterol-sensing domain reminiscent of that of SCAP, are embedded in the ER membrane<sup>23</sup>. High levels of intermediate lipid species in cholesterol synthesis, such as lanosterol, trigger the binding of HMG-CoA reductase to INSIG, which is also bound constitutively to an ubiquitination complex formed by VCP, GP78 and UBC7. This interaction promotes the ubiquitin-mediated degradation of HMG-CoA reductase<sup>24</sup> (Fig. 1c). As mentioned, HMG-CoA reductase catalyses an early (and rate limiting) step in cholesterol synthesis, but the levels of HMG-CoA reductase are regulated by a slow, transcriptional mechanism that is shut off only after cholesterol levels have been replenished. Hence, the interaction of HMG-CoA reductase with INSIG, leading to its turnover by the proteasome, constitutes a faster regulatory loop that aims to put a brake on cholesterol synthesis when the presence of precursor molecules guarantees its imminent increase.

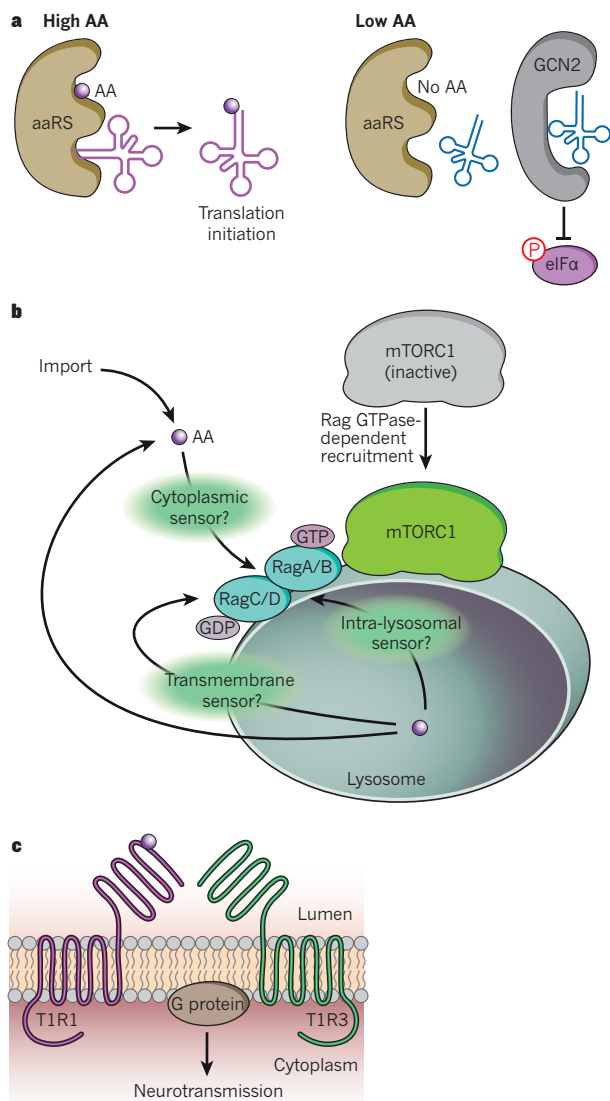
## Sensors upstream of adipokines

Adipokines, hormones secreted by adipocytes, exert systemic effects that include the regulation of appetite, energy expenditure and other processes that contribute to nutrient homeostasis. Their levels do not necessarily reflect circulating lipid levels, but relate to organismal lipid storage<sup>25</sup>, and some adipokines, such as leptin, can be considered as surrogate indicators of lipid-storage abundance. Surprisingly, the identity of the sensor that connects high levels of stored lipids with leptin production remains a mystery, despite the identification of regulatory elements in the promoter region of the *LEP* gene<sup>26</sup>. We know considerably more regarding the systemic effects downstream



**Figure 1 | Lipid-sensing mechanism.** **a**, Fatty-acid (FA) detection mechanisms by GPR40 and GPR120 (left) and CD36 (right). GPR family members are expressed in several cell types including enteroendocrine cells, taste buds and white adipocytes; in the enteroendocrine cells, binding to fatty acids occurs on the luminal side, and the signal is transduced through a G protein, leading to the release of incretins into the circulation; in taste buds, they trigger the release of neurotransmitters; and in white adipocytes, activation of GPR120 indirectly promotes glucose uptake. Binding of CD36 to free fatty acids in oral taste buds triggers calcium release from the endoplasmic reticulum (ER) and neurotransmission; in enterocytes, it directly promotes fatty-acid uptake. **b**, Cholesterol sensing by SREBP1 cleavage-activating protein (SCAP). In the presence of cholesterol, the SCAP–SREBP complex binds the INSIG proteins at the ER membrane and remains anchored in the ER. When cholesterol is absent and SCAP–SREBP does not bind INSIG, the complex traffics to the Golgi where the cytoplasmic tail of SREBP gets released by proteolytic cleavage, and triggers a cholesterol synthesis transcriptional program at the nucleus, including the synthesis of HMG-CoA reductase (HMGCR). **c**, The enzyme HMGCR catalyses a rate-limiting step in cholesterol synthesis, and is synthesized when cholesterol levels are low. HMGCR is embedded in the ER membrane and has cytoplasmic domains, which include its catalytic activity (right). In the presence of abundant intermediate species in the cholesterol biosynthetic pathway (such as lanosterol), HMGCR interacts with the INSIG proteins, constitutively bound to an ubiquitination complex. This leads to HMGCR ubiquitination and degradation and halts the synthesis of cholesterol in a rapid regulatory mechanism, which is key to the anticipation of an imminent increase in cholesterol levels.

of leptin. Leptin receptor (LEPR) is expressed both in the central nervous system and in peripheral tissues and its activation coordinates food intake and organismal metabolism. In hypothalamic neurons that suppress appetite (anorexigenic neurons), leptin activity



**Figure 2 | Amino-acid-sensing mechanisms.** **a**, GCN2 detects insufficiencies of cellular amino acids (AAs). During low levels of any amino acid, its cognate aminoacyl transfer RNA synthetase (aaRS) fails to load the tRNA (purple and blue structures). The unloaded tRNA is then detected by GCN2 kinase, which halts translation initiation. **b**, Mechanistic target of rapamycin complex 1 (mTORC1) is activated downstream of elevated intracellular amino acids through its recruitment to the outer lysosomal surface by a Rag GTPase-mediated mechanism. The identity of the sensor for amino acids remains unclear, and several non-mutually exclusive possibilities exist: an intra-lysosomal sensor that transduces the signal through the membrane, a lysosomal transmembrane sensor that both detects and transduces the signal, and a cytoplasmic sensor that operates downstream of amino-acid export from the lysosome. **c**, Sensing of extra-organismal amino acids by oral taste receptors. The heterodimeric receptor T1R1–T1R3 binds amino acids at high concentrations only, and triggers a signal transduction cascade through a G protein. In the intestinal epithelium, it also leads to the localization of GLUT2 to the apical membrane, facilitating glucose import.

antagonizes the effect of appetite-stimulating neuropeptides and neurotransmitters. Lipid mobilization by adipocytes, as occurs in fasting states, results in decreased leptin production, thereby stimulating appetite and promoting nutrient acquisition behaviour. Indeed, mutations in the *LEPR* gene were found in people who are morbidly obese<sup>27</sup>, and mice harbouring inactivating mutations in *Lep*<sup>28</sup> or *Lepr*<sup>29</sup> are hyperphagic to the extent that they can be double the mass of normal mice.

In addition to leptin, adipocytes also synthesize the hormone

adiponectin (encoded by *ADIPOQ*)<sup>30,31</sup>, although we have even less of an understanding of the regulation of its production<sup>32</sup>. In contrast to leptin, circulating adiponectin levels inversely correlate with lipid storage, and this adipokine exerts a multitude of systemic effects that include the promotion of energy expenditure, insulin sensitivity and loss of appetite<sup>33–35</sup>. Mutations and polymorphisms in the human *ADIPOQ* gene strongly correlate with obesity and the development of type 2 diabetes<sup>36–38</sup>.

### Amino-acid sensing

Amino acids are the building blocks for proteins, the most abundant macromolecules in cells. Protein synthesis is energetically expensive and complex; accordingly, cells sense extracellular and intracellular amino acids to couple their abundance to use. When amino acids are scarce, proteins constitute reservoirs of amino acids that catabolic programs, such as proteasome-mediated degradation and autophagy, mobilize. Amino acids are subsequently recycled and allocated for the synthesis of specific proteins required under nutrient limitation. Furthermore, during periods of prolonged starvation and hypoglycaemia, amino acids are catabolized for the production of other forms of energy, such as glucose and ketone bodies, which are required to fuel the particular needs of certain organs (for example, the brain). Hence, the accurate sensing of amino-acid levels is key for the efficient regulation of protein and amino-acid synthesis and catabolism, as well as for the control of food intake.

#### GCN2

In protein synthesis, no amino acid compensates for the absence of another, therefore, the cell must be able to efficiently detect the lack of any amino acid to prevent potential failures in peptide-chain synthesis. The structural unit of protein-synthesis machinery, the ribosome, incorporates amino acids into a nascent peptide by the sequential binding of a specific transfer RNA covalently linked to its cognate amino acid. Amino-acid-specific aminoacyl tRNA synthetases (aaRSs) execute the loading of amino acids to their cognate tRNAs<sup>39</sup>, and uncharged tRNAs accumulate during low levels of free amino acids. Failure to finish a peptide chain due to a stalled ribosome under amino-acid scarcity is inefficient and energetically onerous, so cells anticipate this situation by preventing the initiation of translation. The mechanism involves a single protein that is able to detect any uncharged tRNA, regardless of its amino-acid specificity, allowing for the detection of low levels of any amino acid in the context of an abundance of the other 19 amino acids. This protein is general control nonderepressible 2 (GCN2), which has a high affinity to all uncharged tRNAs<sup>40</sup> (Fig. 2a), and represents an elegant example of amino-acid sensing by the detection of a surrogate molecule. Under low intracellular amino acid levels, the binding of GCN2 to a given uncharged tRNA triggers a conformational change that leads to kinase activation and inhibitory phosphorylation of a key early activator of translation initiation: eukaryotic translation initiator factor 2α (eIF2α)<sup>41</sup>. Mouse models have proven the importance of GCN2 and eIF2α in mammalian responses to transient drops in amino acids<sup>42,43</sup> and, interestingly, this amino-acid-sensing pathway seems to play a key part in the central nervous system for the detection of imbalances in amino-acid composition in food, independent of taste<sup>44–46</sup>.

Inhibition of protein synthesis by GCN2 and eIF2α occurs in concert with other cellular responses to amino-acid depletion, such as the inhibition of the mechanistic target of rapamycin (mTOR) pathway (see 'mTORC1'). This restricts translation to those messenger RNAs encoding proteins required for cellular adaptation to nutrient starvation, while impairing synthesis of most other proteins<sup>47</sup>. Minimizing translation also enables amino acids to be used as energy sources.

#### mTORC1

The mTOR kinase, when part of mTOR complex 1 (mTORC1), controls cellular energetics by inducing numerous anabolic processes, including protein and lipid synthesis<sup>48</sup>. Growth factors activate mTORC1 through



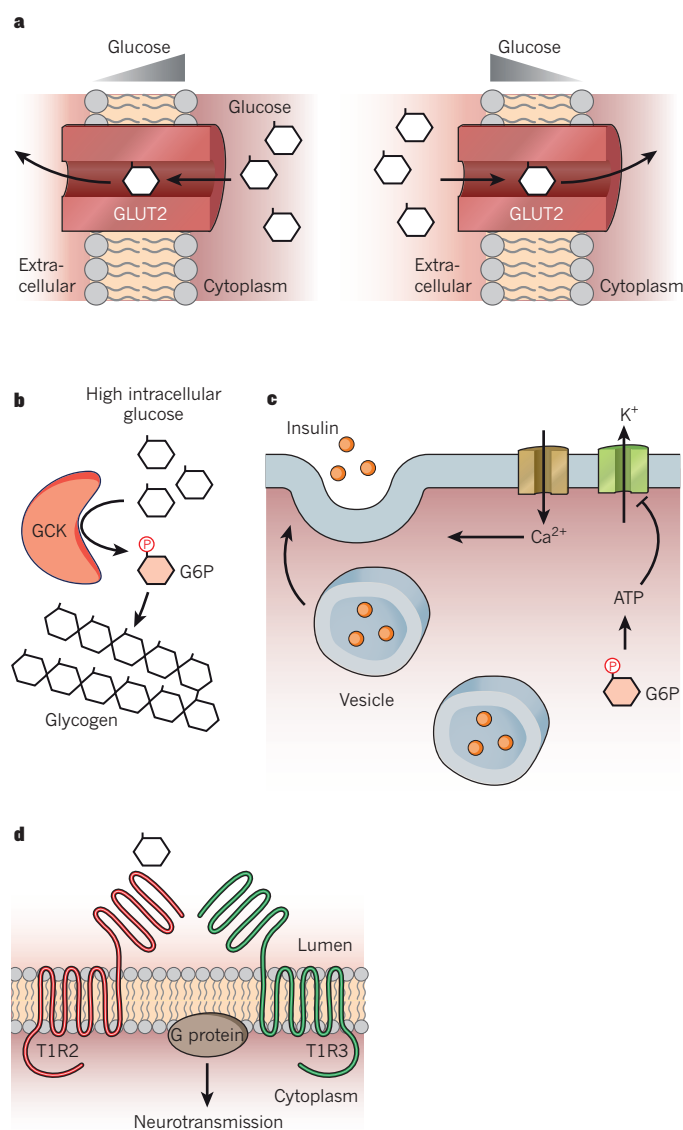
a well-understood signal transduction cascade initiated by the binding of a receptor at the plasma membrane, and culminating in the activation of the Rheb GTPase. Rheb directly binds mTORC1 and activates its kinase in a growth-factor-dependent manner<sup>49–52</sup>. In addition to regulation by hormones, intracellular amino acids also activate mTORC1, so the complex integrates information on both systemic and cellular nutrient levels. In spite of the fact that mTORC1 activity is highly responsive to changes in amino-acid levels, it is not an amino-acid sensor. Indeed, mTORC1 activation is one of the several examples of a key sensing signalling process for which, despite intense interest, the actual nutrient sensors remain unidentified (Fig. 2b). mTORC1 is not equally sensitive to all amino acids — leucine, for example, is particularly important for its activation<sup>53</sup>. We can only speculate about the selective importance of leucine levels for mTORC1 activation; it is one of the most abundant amino acids in proteins, and hence, more likely to be limiting during protein synthesis. Intriguingly, GCN2-knockout mice fed a leucine-deficient diet have a more severe phenotype than the same animals fed diets lacking tryptophan or glycine<sup>43</sup>. Thus, leucine seems to be crucial for the organismal sensing of amino-acid sufficiency and deprivation by different pathways. The molecular characterization of the amino-acid-dependent activation of mTORC1 started only a few years ago with the identification of the Rag family of GTPases<sup>54,55</sup>, which regulate mTORC1 through a mechanism distinct to that of growth factors. Whereas growth factors regulate the kinase activity of mTORC1, the Rag GTPases recruit mTORC1 to the outer lysosomal surface, an essential step in its activation<sup>56</sup>. Because mTORC1 kinase activation by Rheb occurs at the outer lysosomal surface, it is only possible following Rag GTPase-dependent recruitment of mTORC1 (Fig. 2b). Hence, amino-acid abundance and the consequent recruitment of mTORC1 is a prerequisite for the activation of mTORC1 by growth factors (Fig. 2). Although the sensors for amino acids have not been identified so far, a few pieces in the puzzle of amino-acid-dependent regulation of mTORC1 have recently been added. Cell-based biochemical studies have identified the proteins responsible for tethering the Rag proteins to the lysosomal surface<sup>56</sup>, guanine exchange factors (GEFs) and GTPase-activating proteins (GAPs), as well as other regulatory proteins operating upstream of the Rag GTPases<sup>57–63</sup>.

The reason for a lysosomal-centred mechanism of mTORC1 activation may be puzzling, but independent pieces of evidence suggest that the lysosome has a key role in amino-acid homeostasis. The yeast vacuole, an organelle equivalent to the mammalian lysosome, accumulates nutrients such as amino acids<sup>64</sup>, and mTORC1 recruitment is conserved in yeast<sup>65</sup>. In addition, high intraluminal concentrations of certain amino acids have also been shown in lysosomes<sup>66</sup>. Protists such as *Dictyostelium discoideum* obtain energy through phagocytosis and lysosomal degradation<sup>67</sup>, which is followed by a transient increase in intralysosomal nutrient levels. Finally, both the lysosome and the vacuole are the organelles in which amino acids and other nutrients are scavenged from cellular components, through the catabolic process of autophagy. Hence, high levels of amino acids within the lysosome or vacuole system seem to reflect, to some extent, cellular amino-acid abundance, and so it is reasonable to couple its sensing with recruitment and activation of mTORC1 — a crucial regulator of most anabolic processes, including protein synthesis.

Germline and sporadic mutations in genes involved in the signal transduction of nutrient levels upstream of the Rag GTPases have been found in human syndromes characterized by growth defects, neurological disorders, skin and immunological problems, and tumorigenesis<sup>60,61,68–70</sup>.

### Amino-acid-sensing taste receptors

As strict heterotrophs, mammals must obtain energy and nutrients from external organic sources. Predicting the nutritional value of food before digestion allows for the accurate selection of food sources and for the anticipation of increased nutrient abundance. Several mechanisms act synergistically, including experience and social rules in humans, but a fundamental nutrient-sensing event occurs at the level of the oral taste buds. Nutrient sensing by taste receptors is not just a means of sensing extracellular nutrients, it is



**Figure 3 | Glucose-sensing mechanisms.** **a**, Glucose sensing by the GLUT2 transporter. Owing to low affinity, this transporter actively imports glucose only during high glycaemic states (right). Its bidirectional properties mean it can also export glucose from hepatocytes into the circulation under hypoglycaemic states if hepatic gluconeogenesis and glycogen breakdown raise the intrahepatic glucose levels (left). **b**, Intracellular glucose sensing by glucokinase (GCK) in hepatic and pancreatic cells. GCK has low affinity for glucose, and shunts glucose-6-phosphate (G6P) into either glycolysis or glycogen synthesis only when glucose is abundant. **c**, The mechanism of insulin release downstream of glucose sensing in pancreatic  $\beta$ -cells. This is a multi-step process that relies on glucose phosphorylation by GCK, subsequent ATP production and ATP-mediated blockade of potassium channels. This leads to a calcium influx that facilitates the release of insulin from vesicles into the bloodstream. **d**, Extra-organismal glucose sensing by oral taste receptors. The dimeric receptors T1R2–T1R3 bind at high concentrations of glucose, sucrose, fructose and artificial sweeteners only, and trigger a signal transduction cascade through a G protein.

a mechanism of extra-organismal sensing that allows the interrogation of prospective food sources. In humans, taste is divided into five categories: sweet, umami, bitter, sour and salty, and is generated by signals elicited in taste buds, groups of cells in the tongue, palatal and oesophageal epithelium. Within these cells, the taste receptors are, logically, exposed in the apical membrane oriented towards the environment<sup>71</sup>.

Taste receptors belong to the T1R and T2R families of G-protein-coupled receptors, and are characterized by seven transmembrane

domains with an extracellular N-terminus and an intracellular C-terminus. (For molecular and genetic information regarding the different members of the taste receptor genes see ref. 71.) The T2R family is involved in the detection of bitter molecules, a category that includes potentially toxic compounds, and two T1R family members are responsible for sensing the presence of amino acids (the umami taste). Although other taste receptors also exist<sup>71,72</sup>, elegant genetic studies using heterologous expression experiments showed that the T1R1–T1R3 heterodimer senses amino acids (Fig. 2c). Human amino-acid taste receptors have a particularly high affinity to glutamate, but other L-amino acids also serve as ligands, whereas D-amino acids do not<sup>73</sup>. Amino-acid binding to a taste receptor triggers signal transduction through the plasma membrane, followed by G-protein activation and neurotransmitter release<sup>74</sup>, which is then integrated with other neurotransmission events at the level of the central nervous system.

In addition to the presence of taste buds in the oral epithelium, taste receptors also exist in endocrine cells in certain regions of the gut<sup>75</sup>. Intestinal taste receptors operate through G-protein activation in a similar manner to that of the oral epithelium, but instead of inducing the release of a neurotransmitter that activates an afferent signal to the brain, the cascade elicited by enteral taste receptors culminates in the release of incretins into the blood circulation, serving as an anticipatory signal for the imminent digestion of, and systemic increase in, nutrients.

Interestingly, extracellular amino-acid sensing at the plasma membrane by taste receptors can modulate mTORC1 activation without

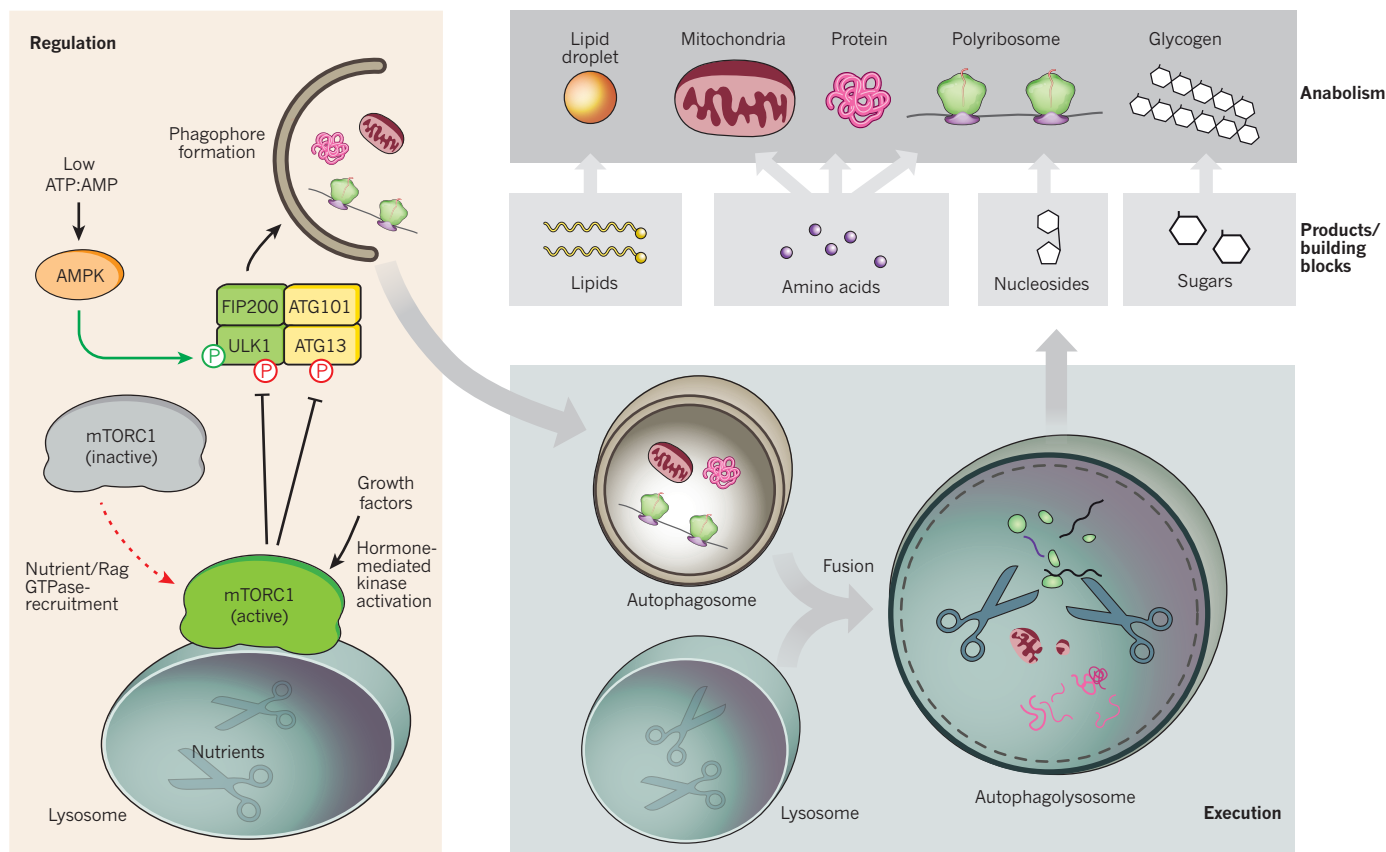
affecting intracellular amino-acid levels<sup>76</sup>, a meaningful cross-talk that engages the anabolic machinery of the cell in anticipation of an elevation in intracellular amino-acid levels, following import.

### Glucose sensing

Mammals rely on multiple ways of maintaining glucose levels within a narrow physiological range. Glucose intake, storage, mobilization and breakdown are tightly regulated at different levels, and multiple mechanisms of glucose sensing coexist: extra-organismal, extracellular and intracellular. In addition, a network of hormone signals, exemplified by insulin and glucagon, aim to coordinate coherent responses to systemic glucose levels in distant organs. Deregulated glucose homeostasis mechanisms, from glucose sensing to import, storage and mobilization underlie the pathogenesis of human diseases such as type 2 diabetes.

### Glucokinase

Glucokinase (GCK) catalyses the first step in the storage and consumption of glucose, glycogen synthesis and glycolysis, and its function constitutes a simple, direct intracellular nutrient-sensing mechanism that controls systemic glucose homeostasis. Like all hexokinases, GCK phosphorylates glucose to make glucose-6-phosphate (G6P), but unlike the other isozymes, only GCK functions as a glucose sensor<sup>77</sup>. This uniqueness occurs because unlike the other hexokinases, which have  $K_m$  values (an inverse measure of affinity) for glucose much below the minimum physiological level of glucose, GCK has a significantly lower affinity and is only active when glucose



**Figure 4 | Nutrients and autophagy.** Autophagy serves as an internal source of stored nutrients under conditions of nutrient limitation. Two main regulatory inputs for autophagy are AMP-activated protein kinase (AMPK) and mechanistic target of rapamycin complex 1 (mTORC1). Autophagy initiation can be promoted by the activation of ULK1 through AMPK-dependent phosphorylation during low ATP:AMP ratios. mTORC1 is activated by growth factor-mediated signal transduction at the outer lysosomal surface if cellular amino acids and glucose have recruited mTORC1 through the action of the Rag

GTPases. Activated mTORC1 inhibits ULK1 and ATG13 by phosphorylation. Hence, low nutrients promote autophagy by the inhibition of mTORC1. Autophagy starts with the engulfment of the cellular constituents glycogen, lipids from lipid droplets, soluble proteins, ribosomes or organelles in a double-membrane structure that then fuses with lysosomes, in which enzymatic breakdown occurs. The products of autophagy, basic nutrients (sugars, lipids, amino acid and nucleosides), are then exported into the cytoplasm, in which they may be used as a source of energy, or re-used for anabolism.

levels are relatively high (around 120 mg dl<sup>-1</sup>, or 7 mM, or greater). Hence, although the other hexokinases function as ‘phosphorylation machines’ regardless of the actual glucose levels, GCK is active only during glucose abundance, and it controls systemic glucose fate through its effects in the liver and pancreas (Fig. 3b). The liver maintains glycaemia through gluconeogenesis and glycogen breakdown during periods of systemic glucose scarcity, or by storing glucose in the form of glycogen when it is in excess<sup>78</sup>. GCK is the most abundant hexokinase in the liver and because it is inactive under conditions of glucose limitation, it permits export of unphosphorylated glucose from the liver in order to supply the energetic demands of the brain and muscles. When hepatic glucose levels are high, GCK-mediated conversion of glucose to the metabolic intermediate G6P allows it to be shunted into glycolysis (for energy production) or glycogen synthesis (for storage).

GCK is also expressed in  $\beta$ -cells (see ‘GLUT2’), and in neurons and glial cells in the hypothalamus. Although work remains to be done to understand the role of this glucose sensor in the brain, systemic effects, such as feeding responses and insulin release, are likely to be downstream of hypothalamic GCK activity<sup>79</sup>. Dozens of germline mutations in GCK in people with abnormal glycaemia and diabetes<sup>80</sup>, together with conditional deletion of the murine *Gck* gene in the liver and pancreas<sup>81</sup>, support the fundamental role of GCK in maintaining organismal glucose homeostasis.

## GLUT2

The glucose transporter GLUT2 (also known as SLC2A2) is a sensor of extracellular glucose levels, and like GCK, GLUT2 has a higher  $K_m$  (20 mM) than other glucose transporters of the same family. The  $K_m$  for GLUT1 is around 1 mM and that of GLUT4 is about 5 mM (ref. 82), so they are close to saturation even during fasting glycaemia (around 4 mM). The low affinity of GLUT2, by contrast, allows for efficient transport of glucose across the plasma membrane only when glycaemia is high, but not under the low concentrations that still saturate the other transporters. Accordingly, GLUT2 has crucial roles in directing organismal glucose handling after feeding. Hepatic glucose import mediated by GLUT2 is followed by GCK-dependent phosphorylation for storage and energy production, as already described. Importantly, during periods of low glycaemia, hepatic glycogenolysis and gluconeogenesis increase intrahepatic glucose levels. Because GLUT2 can transport glucose in a bidirectional manner, it exports glucose to the circulation (Fig. 3a). Hence, GLUT2-mediated import occurs only during transient hyperglycaemic states, and GLUT2-mediated export only happens when intrahepatic glucose levels are high, thus constituting a key controller of glucose homeostasis. Not surprisingly, inactivating mutations in GLUT2 lead to human metabolic disorders, such as Fanconi–Bickel syndrome characterized by deregulated glycogen accumulation, hepatomegaly and hypoglycaemia, among other symptoms of disrupted glycaemic homeostasis<sup>83</sup>.

$\beta$ -Cells in the pancreas have a specialized role in sensing systemic glucose levels, and are responsible for the synthesis and secretion of insulin. Glucose is imported in  $\beta$ -cells and phosphorylated by the tandem of GLUT2 (or GLUT1) and GCK, and, as it is consumed, leads to an increased ATP:ADP ratio. This closes potassium channels at the plasma membrane, and causes the membrane to depolarize. Dissipation of membrane potential results in a transient increase of intracellular calcium that facilitates the fusion of insulin-containing vesicles with the plasma membrane, releasing its cargo into systemic circulation (Fig. 3c). It is important to mention that whereas the predominant transporter in murine  $\beta$ -cells is GLUT2, the relative abundance of the GLUT2 transporter in human islets seems to be minor compared with that of the high affinity GLUT1 transporter — so the relevance of GLUT2 for glucose transport in human  $\beta$ -cells is not clear<sup>84</sup>.

Elevated sugar intake and chronic hyperglycaemia deregulate normal glucose sensing through several mechanisms, including ER stress, increased intracellular Ca<sup>2+</sup> levels, mitochondrial dysfunction,

reactive oxygen species and chronic inflammation, all of which contribute to the corruption of insulin secretion in type 2 diabetes<sup>85</sup>.

Finally, although the other glucose transporters (such as GLUT1 and GLUT4) do not behave as sensors, their activities and effects are regulated by different means in order to meet particular requirements of glucose use and storage. GLUT4 is expressed in skeletal muscle and adipose tissue, two organs important for post-prandial glucose uptake and storage<sup>82</sup>, and although GLUT4 has a low  $K_m$ , glucose uptake in these organs is a regulated process. Insulin triggers a PI(3)K–AKT dependent signal transduction cascade that results in GLUT4 localization to the plasma membrane, allowing glucose uptake in these tissues<sup>86</sup>. Because glucose import and storage are insulin dependent, and thus secondary to direct glucose-sensing mechanisms in the liver and pancreas, they occur only after the organism has reached a threshold of internal glucose abundance. GLUT1 is expressed in fetal tissues and its constant activity provides glucose to all tissues to sustain the rapid growth of the organism.

## AMPK and ATP:AMP ratios

AMP-activated protein kinase (AMPK) is a fundamental regulator of cellular metabolism and coordinates several metabolic responses in different cell types. It is exquisitely responsive to cellular energy levels; as a surrogate sensing mechanism for glucose abundance, increased levels of AMP and ADP directly activate the kinase. AMPK has been the subject of a number of excellent reviews addressing its activation, regulation, and downstream consequences<sup>87,88</sup>, and will be briefly discussed here in the context of the regulation of autophagy.

## mTORC1 and the sensing of glucose

The regulation of mTORC1 through Rag-GTPase-mediated recruitment is not restricted to amino acids; cellular glucose levels also affect the activity of the Rag GTPases<sup>89</sup>. In contrast to the identification of some molecular players involved in the activation of the Rag GTPases by cellular amino acids, less clear is the mechanism by which glucose regulates the Rag GTPases. Some aspects downstream of glucose and amino-acid sensing are shared, such as the involvement of the lysosomal v-ATPase<sup>48,89,90</sup>, but additional players remain unidentified. Because the amino-acid- and glucose-sensing mechanisms are generally independent phenomena, as we illustrate, it is very likely that amino-acid and glucose sensing upstream of mTORC1 occur in parallel and converge upstream of the Rag GTPases, but precisely how this integration occurs is unresolved.

## Glucose-sensing taste receptors

In a similar manner to amino-acid sensing in taste buds by T1R1–T1R3, the heterodimer composed of T1R2–T1R3 constitutes the glucose taste receptor (Fig. 3d). The extracellular N-terminal domains of both T1R1 and T1R2 are essential for determining their specificity for their natural ligands<sup>91</sup>. Millimolar concentration of the saccharides glucose, fructose or sucrose activate the T1R2–T1R3 receptor<sup>92</sup>; this concentration may seem high, but sucrose concentration in an apple is around 100–200 mM, and so this process is selective and efficient for the detection of highly energetic foods.

T1R2–T1R3 receptors are also expressed in the intestinal epithelium, and although the sensing process is identical to that of the oral epithelium, the signal transduction does not trigger an afferent signal to the brain, but results in the transient localization of the GLUT2 transporter to the apical membrane, leading to increased absorption of glucose from the intestinal lumen after feeding<sup>93,94</sup>.

In addition to natural ligands, glucose taste buds also respond to artificial sweeteners such as saccharine, cyclamate and aspartame<sup>92</sup>. Activation of glucose taste receptors by artificial ligands has clinical implications for obesity and type 2 diabetes, as sweeteners may increase nutrient absorption and activate other nutrient-sensing signalling cascades at different levels, regardless of nutritional value. Indeed, some studies have shown that consumers of artificial



sweeteners are at higher risk of developing metabolic disease<sup>95</sup>. The phenomenon of artificial activation of this nutrient-sensing pathway is currently an active field of research.

### Accessing internal nutrient stores through autophagy

Because environmental nutrient availability can be intermittent, cells and organisms have evolved efficient ways of storing nutrients during periods of abundance. This occurs in unicellular organisms and is more obvious and prominent in animals, with the emergence of organs specialized in nutrient storage, such as fat tissue, the liver and skeletal muscle. Mammalian cells accumulate and store glucose in the form of glycogen, lipids within lipid droplets and internal membranes, and amino acids in proteins and organelles; all of which can be mobilized and catabolized to endure periods of nutrient limitation. Cells exploit different means to obtain the basic nutrients from internal stores, including autophagy, the controlled process of recycling of cellular constituents confined within a double membrane structure. Autophagy starts with the *de novo* formation of a membrane structure termed the phagophore, which engulfs its cargo and closes as a cytoplasmic double-membrane autophagosome. An autophagosome then fuses with a lysosome, which leads to the enzymatic breakdown of the autophagosomal cargo into its basic building blocks, which are then exported from the autophagolysosomes and further catabolized to produce energy, or used again in other anabolic reactions (Fig. 4).

The process of autophagy is unique because it can target any cellular component and nutrient storage depot, and, as a key internal source when nutrients are scarce, is highly regulated at multiple levels by nutrients and nutrient signalling<sup>96</sup>. AMPK, directly activated by a low ATP:ADP ratio, phosphorylates and activates ULK1, a kinase that regulates autophagy initiation<sup>97,98</sup>. AMPK also activates the FOXO transcription factors, which transactivate the ATG genes responsible for the initiation and completion of autophagy<sup>99</sup>. Hence, AMPK acutely regulates autophagy, as well as by means of a slower, transcriptional mechanism.

A crucial regulator of autophagy, as shown in all eukaryotes using both cultured cells and model organisms, is mTORC1, through its inhibitory phosphorylation of ULK1 and ATG13 (ref. 100). mTORC1 seems to play a dominant part in the regulation of autophagy, as mTORC1 inhibition is sufficient to induce it<sup>101</sup>, whereas its constitutive activation is sufficient to block it<sup>89</sup>. Nutrient depletion is perhaps the most potent inducer of autophagy, and the regulation of mTORC1 by the Rag GTPases downstream of nutrient scarcity seems to be essential for the regulation of autophagy. Mice with constitutive RagA activity, and hence, constitutive activation of mTORC1 regardless of nutrient levels, develop normally but die within the first day of life, similar to mice lacking the essential autophagy genes *Atg5* and *Atg7* (refs 89,102,103). Constitutive RagA activity in neonatal mice leads to a profound glucose and amino-acid homeostasis defect secondary to an impairment in the detection of nutrient shortage after the transplacental supply of nutrients is interrupted at birth. This leads to constitutive mTORC1 activity and the consequent inability to trigger autophagy.

In addition to the regulation of autophagy initiation, mTORC1 activity is required for autophagy termination<sup>104</sup>. Cellular free amino acids, produced by autophagy, result in an increase in mTORC1 activity and the reformation of lysosomes. Systemic levels of nutrients also regulate autophagy through the effects of insulin<sup>105</sup>. The intracellular cascade of insulin activates AKT, a positive input for mTORC1, and a negative regulator of the FOXO transcription factors. Hence, both local and systemic nutrients regulate the process. In addition to nutrients, stresses such as hypoxia, ER stress and DNA damage also regulate autophagy<sup>106</sup>.

Several studies that generated autophagy-deficient tissues in a temporal specific manner have determined the importance of autophagy in mammalian physiology. Besides the aforementioned role of autophagy in the early neonatal starvation period<sup>102,103</sup>, autophagy is

essential for the survival of embryos in the pre-implantation stage<sup>107</sup>. Whole-body acute deletion of autophagy genes in adult mice eventually culminates in neurodegeneration and death, presumably owing to the accumulation of harmful organelles and proteins, which probably cause neuronal toxicity<sup>108,109</sup>. Liver-specific impairment in autophagy results in accumulation of abnormal cellular endomembranes, mitochondria and ubiquitinated proteins<sup>103</sup>, and impaired lipid mobilization<sup>110</sup>. Impaired autophagy seems to preferentially affect cells specialized in vesicle trafficking, such as lymphocytes and  $\beta$ -cells<sup>111</sup>, but some of these effects may be due to a deranged endomembrane trafficking system, rather than a direct consequence of a nutrient homeostasis defect.

### Future directions

Despite intense research, our understanding of nutrient-sensing mechanisms is far from complete. For instance, we have not yet deciphered what links lipid storage levels with leptin synthesis and release. Equally unclear is what the glucose and amino-acid sensors upstream of mTORC1 are. Towards the identification of nutrient sensors upstream of mTORC1, the lysosome seems to be a key organelle in sensing; however, we still need to determine what is sensed, and how, at the lysosome. Besides these and other fundamental unanswered questions of direct nutrient sensing, the mechanisms discussed in this Review are outlined mostly in a modular manner. This reflects our lack of an integrative view of the nutrient-sensing pathways, and connecting the different aspects of nutrient sensing will be one of the challenges of future research. We know that mTORC1 is a node at which hormone and nutrient inputs converge, but we still do not know whether these signalling cascades cross-talk upstream of mTORC1. A complete view of nutrient-sensing mechanisms will address potential cross-regulation between different nutrient-sensing pathways, but also incorporate regulation by other signalling events. For example, we know some of the consequences of chronic inflammation in deregulating nutrient-sensing mechanisms and the signalling cascades downstream, such as those that occur in obese states, but how exercise modulates nutrient inputs, or how ageing affects nutrient-sensing abilities, remain to be determined. From an experimental point of view, advances in genomics will probably provide insight into clinical conditions secondary to deregulated nutrient sensing, such as the identification of novel mutations and polymorphisms in humans. Finally, nutrient abundance not only affects the onset of diabetes, but also influences cancer development and the ageing process. Nutrient sensing and metabolism in cancer cells has received a new wave of attention, partly thanks to advances in next-generation sequencing and metabolomics. On the one hand, cancer cells are exposed to limited nutrients owing to poor vasculature, and deregulated proliferation poses energetic and nutrient demands and liabilities, which act in concert with aberrant activation of growth signals. On the other hand, one of the most successful interventions against the onset of ageing is limitation of nutrient intake, or caloric restriction<sup>112</sup>. Hence, understanding normal nutrient-sensing mechanisms is a prerequisite for designing better interventions against human disease beyond diabetes. ■

Received 1 October; accepted 2 December 2014.

1. Wu, G. & Morris, S. M. Arginine metabolism: nitric oxide and beyond. *Biochem. J.* **336**, 1–17 (1998).
2. Reeds, P. J. Dispensable and indispensable amino acids for humans. *J. Nutr.* **130**, 1835S–1840S (2000).
3. Richieri, G. V. & Kleinfeld, A. M. Unbound free fatty acid levels in human serum. *J. Lipid Res.* **36**, 229–240 (1995).
4. Itoh, Y. *et al.* Free fatty acids regulate insulin secretion from pancreatic  $\beta$  cells through GPR40. *Nature* **422**, 173–176 (2003).
5. Hirasawa, A. *et al.* Free fatty acids regulate gut incretin glucagon-like peptide-1 secretion through GPR120. *Nature Med.* **11**, 90–94 (2005).
6. Oh, D. Y. *et al.* GPR120 is an omega-3 fatty acid receptor mediating potent anti-inflammatory and insulin-sensitizing effects. *Cell* **142**, 687–698 (2010).
7. Ichimura, A. *et al.* Dysfunction of lipid sensor GPR120 leads to obesity in both mouse and human. *Nature* **483**, 350–354 (2012).

8. Oh, D. Y. *et al.* A Gpr120-selective agonist improves insulin resistance and chronic inflammation in obese mice. *Nature Med.* **20**, 942–947 (2014).
9. Pepino, M. Y., Kuda, O., Samovski, D. & Abumrad, N. A. Structure-function of CD36 and importance of fatty acid signal transduction in fat metabolism. *Annu. Rev. Nutr.* **34**, 281–303 (2014).
10. Laugerette, F. *et al.* CD36 involvement in orosensory detection of dietary lipids, spontaneous fat preference, and digestive secretions. *J. Clin. Invest.* **115**, 3177–3184 (2005).
11. Cartoni, C. *et al.* Taste preference for fatty acids is mediated by GPR40 and GPR120. *J. Neurosci.* **30**, 8376–8382 (2010).
12. Martin, C. *et al.* The lipid-sensor candidates CD36 and GPR120 are differentially regulated by dietary lipids in mouse taste buds: impact on spontaneous fat preference. *PLoS ONE* **6**, e24014 (2011).
13. Pepino, M. Y., Love-Gregory, L., Klein, S. & Abumrad, N. A. The fatty acid translocase gene CD36 and lingual lipase influence oral sensitivity to fat in obese subjects. *J. Lipid Res.* **53**, 561–566 (2012).
14. Brown, M. S. & Goldstein, J. L. A receptor-mediated pathway for cholesterol homeostasis. *Science* **232**, 34–47 (1986).
15. Brown, A. J., Sun, L., Feramisco, J. D., Brown, M. S. & Goldstein, J. L. Cholesterol addition to ER membranes alters conformation of SCAP, the SREBP escort protein that regulates cholesterol metabolism. *Mol. Cell* **10**, 237–245 (2002).  
**This paper demonstrates the functional regulation of SCAP-protein conformation by cholesterol levels within the ER membrane, providing strong support for its cholesterol-sensing ability.**
16. Radhakrishnan, A., Sun, L.-P., Kwon, H. J., Brown, M. S. & Goldstein, J. L. Direct binding of cholesterol to the purified membrane region of SCAP: mechanism for a sterol-sensing domain. *Mol. Cell* **15**, 259–268 (2004).
17. Feramisco, J. D. *et al.* Intramembrane aspartic acid in SCAP protein governs cholesterol-induced conformational change. *Proc. Natl Acad. Sci. USA* **102**, 3242–3247 (2005).
18. Yang, T. *et al.* Crucial step in cholesterol homeostasis: sterols promote binding of SCAP to INSIG-1, a membrane protein that facilitates retention of SREBPs in ER. *Cell* **110**, 489–500 (2002).
19. Radhakrishnan, A., Goldstein, J. L., McDonald, J. G. & Brown, M. S. Switch-like control of SREBP-2 transport triggered by small changes in ER cholesterol: a delicate balance. *Cell Metab.* **8**, 512–521 (2008).
20. Motamed, M. *et al.* Identification of luminal Loop 1 of Scap protein as the sterol sensor that maintains cholesterol homeostasis. *J. Biol. Chem.* **286**, 18002–18012 (2011).
21. Zhang, Y., Motamed, M., Seemann, J., Brown, M. S. & Goldstein, J. L. Point mutation in luminal loop 7 of Scap protein blocks interaction with loop 1 and abolishes movement to Golgi. *J. Biol. Chem.* **288**, 14059–14067 (2013).
22. Jeon, T.-I. & Osborne, T. F. SREBPs: metabolic integrators in physiology and metabolism. *Trends Endocrinol. Metab.* **23**, 65–72 (2012).
23. Sever, N., Yang, T., Brown, M. S., Goldstein, J. L. & DeBose-Boyd, R. A. Accelerated degradation of HMG CoA reductase mediated by binding of insig-1 to its sterol-sensing domain. *Mol. Cell* **11**, 25–33 (2003).
24. Song, B.-L., Sever, N. & DeBose-Boyd, R. A. Gp78, a membrane-anchored ubiquitin ligase, associates with Insig-1 and couples sterol-regulated ubiquitination to degradation of HMG CoA reductase. *Mol. Cell* **19**, 829–840 (2005).
25. Birsoy, K. *et al.* Cellular program controlling the recovery of adipose tissue mass: an *in vivo* imaging approach. *Proc. Natl Acad. Sci. USA* **105**, 12985–12990 (2008).
26. Wrann, C. D. *et al.* FOSL2 promotes leptin gene expression in human and mouse adipocytes. *J. Clin. Invest.* **122**, 1010–1021 (2012).
27. Clément, K. *et al.* A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* **392**, 398–401 (1998).
28. Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425–432 (1994).  
**In this seminal paper, the mouse *Ob* gene and its human homologue *LEP* are identified.**
29. Lee, G. H. *et al.* Abnormal splicing of the leptin receptor in diabetic mice. *Nature* **379**, 632–635 (1996).
30. Scherer, P. E., Williams, S., Fogliano, M., Baldini, G. & Lodish, H. F. A novel serum protein similar to C1q, produced exclusively in adipocytes. *J. Biol. Chem.* **270**, 26746–26749 (1995).
31. Hu, E., Liang, P. & Spiegelman, B. M. AdipoQ is a novel adipose-specific gene dysregulated in obesity. *J. Biol. Chem.* **271**, 10697–10703 (1996).
32. Shehzad, A., Iqbal, W., Shehzad, O. & Lee, Y. S. Adiponectin: regulation of its production and its role in human diseases. *Hormones (Athens)* **11**, 8–20 (2012).
33. Maeda, N. *et al.* Diet-induced insulin resistance in mice lacking adiponectin/ACRP30. *Nature Med.* **8**, 731–737 (2002).
34. Kadowaki, T. *et al.* Adiponectin and adiponectin receptors in insulin resistance, diabetes, and the metabolic syndrome. *J. Clin. Invest.* **116**, 1784–1792 (2006).
35. Waki, H. & Tontonoz, P. Endocrine functions of adipose tissue. *Annu. Rev. Pathol.* **2**, 31–56 (2007).
36. Takahashi, M. *et al.* Genomic structure and mutations in adipose-specific gene, adiponectin. *Int. J. Obes. Relat. Metab. Disord.* **24**, 861–868 (2000).
37. Hara, K. *et al.* Genetic variation in the gene encoding adiponectin is associated with an increased risk of type 2 diabetes in the Japanese population. *Diabetes* **51**, 536–540 (2002).
38. Kondo, H. *et al.* Association of adiponectin mutation with type 2 diabetes: a candidate gene for the insulin resistance syndrome. *Diabetes* **51**, 2325–2328 (2002).
39. Ibba, M. & Soll, D. Aminoacyl-tRNA synthesis. *Annu. Rev. Biochem.* **69**, 617–650 (2000).
40. Dong, J., Qiu, H., Garcia-Barrio, M., Anderson, J. & Hinnebusch, A. G. Uncharged tRNA activates GCN2 by displacing the protein kinase moiety from a bipartite tRNA-binding domain. *Mol. Cell* **6**, 269–279 (2000).
41. Berlanga, J. J., Santoyo, J. & De Haro, C. Characterization of a mammalian homolog of the GCN2 eukaryotic initiation factor 2 $\alpha$  kinase. *Eur. J. Biochem.* **265**, 754–762 (1999).
42. Scheuner, D. *et al.* Translational control is required for the unfolded protein response and *in vivo* glucose homeostasis. *Mol. Cell* **7**, 1165–1176 (2001).
43. Zhang, P. *et al.* The GCN2 eIF2 $\alpha$  kinase is required for adaptation to amino acid deprivation in mice. *Mol. Cell Biol.* **22**, 6681–6688 (2002).
44. Maurin, A.-C. *et al.* The GCN2 kinase biases feeding behavior to maintain amino acid homeostasis in omnivores. *Cell Metab.* **1**, 273–277 (2005).
45. Hao, S. *et al.* Uncharged tRNA and sensing of amino acid deficiency in mammalian piriform cortex. *Science* **307**, 1776–1778 (2005).
46. Guo, F. & Cavener, D. R. The GCN2 eIF2 $\alpha$  kinase regulates fatty-acid homeostasis in the liver during deprivation of an essential amino acid. *Cell Metab.* **5**, 103–114 (2007).
47. Thoreen, C. C. *et al.* A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature* **485**, 109–113 (2012).
48. Laplante, M. & Sabatini, D. M. mTOR signaling in growth control and disease. *Cell* **149**, 274–293 (2012).
49. Garami, A. *et al.* Insulin activation of Rheb, a mediator of mTOR/S6K/4E-BP signaling, is inhibited by TSC1 and 2. *Mol. Cell* **11**, 1457–1466 (2003).
50. Inoki, K., Li, Y., Xu, T. & Guan, K.-L. Rheb GTPase is a direct target of TSC2 GAP activity and regulates mTOR signaling. *Genes Dev.* **17**, 1829–1834 (2003).
51. Tee, A. R., Manning, B. D., Roux, P. P., Cantley, L. C. & Blenis, J. Tuberous sclerosis complex gene products, Tuberin and Hamartin, control mTOR signaling by acting as a GTPase-activating protein complex toward Rheb. *Curr. Biol.* **13**, 1259–1268 (2003).
52. Zhang, Y. *et al.* Rheb is a direct target of the tuberous sclerosis tumour suppressor proteins. *Nature Cell Biol.* **5**, 578–581 (2003).
53. Hara, K. *et al.* Amino acid sufficiency and mTOR regulate p70 S6 kinase and eIF-4E BP1 through a common effector mechanism. *J. Biol. Chem.* **273**, 14484–14494 (1998).  
**This paper explores the amino-acid essentiality for mTORC1 activation, and specific amino-acid requirements independent of the growth-factor-mediated regulation of activity.**
54. Kim, E., Goraksha-Hicks, P., Li, L., Neufeld, T. P. & Guan, K.-L. Regulation of TORC1 by Rag GTPases in nutrient response. *Nature Cell Biol.* **10**, 935–945 (2008).
55. Sancak, Y. *et al.* The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science* **320**, 1496–1501 (2008).  
**References 54 and 55 report the identification of the Rag GTPases as the direct link between amino-acids levels and mTORC1, regulating mTORC1's subcellular localization.**
56. Sancak, Y. *et al.* Ragulator-Rag complex targets mTORC1 to the lysosomal surface and is necessary for its activation by amino acids. *Cell* **141**, 290–303 (2010).
57. Bar-Peled, L., Schweitzer, L. D., Zoncu, R. & Sabatini, D. M. Ragulator is a GEF for the Rag GTPases that signal amino acid levels to mTORC1. *Cell* **150**, 1196–1208 (2012).
58. Bar-Peled, L. *et al.* A tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. *Science* **340**, 1100–1106 (2013).
59. Panchaud, N., Péli-Gulli, M.-P. & De Virgilio, C. Amino acid deprivation inhibits TORC1 through a GTPase-activating protein complex for the Rag family GTPase Gtr1. *Sci. Signal.* **6**, ra42 (2013).
60. Tsun, Z.-Y. *et al.* The folliculin tumor suppressor is a GAP for the RagC/D GTPases that signal amino acid levels to mTORC1. *Mol. Cell* **52**, 495–505 (2013).
61. Petit, C. S., Rocznik-Ferguson, A. & Ferguson, S. M. Recruitment of folliculin to lysosomes supports the amino acid-dependent activation of Rag GTPases. *J. Cell Biol.* **202**, 1107–1122 (2013).
62. Chantranupong, L. *et al.* The sestrins interact with GATOR2 to negatively regulate the amino-acid-sensing pathway upstream of mTORC1. *Cell Rep.* **9**, 1–8 (2014).
63. Peng, M., Yin, N. & Li, M. O. Sestrins function as guanine nucleotide dissociation inhibitors for Rag GTPases to control mTORC1 signaling. *Cell* **159**, 122–133 (2014).
64. Kitamoto, K., Yoshizawa, K., Ohsumi, Y. & Anraku, Y. Dynamic aspects of vacuolar and cytosolic amino acid pools of *Saccharomyces cerevisiae*. *J. Bacteriol.* **170**, 2683–2686 (1988).
65. Binda, M. *et al.* The Vam6 GEF controls TORC1 by activating the EGO complex. *Mol. Cell* **35**, 563–573 (2009).
66. Harms, E., Gochman, N. & Schneider, J. A. Lysosomal pool of free-amino acids. *Biochem. Biophys. Res. Commun.* **99**, 830–836 (1981).
67. Neuhaus, E. M., Almers, W. & Soldati, T. Morphology and dynamics of the endocytic pathway in *Dictyostelium discoideum*. *Mol. Biol. Cell* **13**, 1390–1407 (2002).
68. Lee, J. H. *et al.* *De novo* somatic mutations in components of the PI(3)K-AKT3-mTOR pathway cause hemimegalencephaly. *Nature Genet.* **44**, 941–945 (2012).
69. Bohn, G. *et al.* A novel human primary immunodeficiency syndrome caused by deficiency of the endosomal adaptor protein p14. *Nature Med.* **13**, 38–45 (2007).
70. Efeyan, A., Zoncu, R. & Sabatini, D. M. Amino acids and mTORC1: from lysosomes to disease. *Trends Mol. Med.* **18**, 524–533 (2012).
71. Bachmanov, A. A. & Beauchamp, G. K. Taste receptor genes. *Annu. Rev. Nutr.* **27**, 389–414 (2007).
72. Damak, S. *et al.* Detection of sweet and umami taste in the absence of taste receptor T1r3. *Science* **301**, 850–853 (2003).
73. Nelson, G. *et al.* An amino-acid taste receptor. *Nature* **416**, 199–202 (2002).

74. Chaudhari, N. & Roper, S. D. The cell biology of taste. *J. Cell Biol.* **190**, 285–296 (2010).
75. Wu, S. V. *et al.* Expression of bitter taste receptors of the T2R family in the gastrointestinal tract and enteroendocrine STC-1 cells. *Proc. Natl Acad. Sci. USA* **99**, 2392–2397 (2002).
76. Wauson, E. M. *et al.* The G protein-coupled taste receptor T1R1/T1R3 regulates mTORC1 and autophagy. *Mol. Cell* **47**, 851–862 (2012).
77. Printz, R. L., Magnuson, M. A. & Granner, D. K. Mammalian glucokinase. *Annu. Rev. Nutr.* **13**, 463–496 (1993).
78. Nordlie, R. C., Foster, J. D. & Lange, A. J. Regulation of glucose production by the liver. *Annu. Rev. Nutr.* **19**, 379–406 (1999).
79. Ogunnowo-Bada, E. O., Heeley, N., Brochard, L. & Evans, M. L. Brain glucose sensing, glucokinase and neural control of metabolism and islet function. *Diabetes Obes. Metab.* **16** (Suppl 1), 26–32 (2014).
80. Gloyn, A. L. Glucokinase (GCK) mutations in hyper- and hypoglycemia: maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemia of infancy. *Hum. Mutat.* **22**, 353–362 (2003).
81. Postic, C. *et al.* Dual roles for glucokinase in glucose homeostasis as determined by liver and pancreatic  $\beta$  cell-specific gene knock-outs using Cre recombinase. *J. Biol. Chem.* **274**, 305–315 (1999).
82. Thorens, B. & Mueckler, M. Glucose transporters in the 21st Century. *Am. J. Physiol. Endocrinol. Metab.* **298**, E141–E145 (2010).
83. Santer, R. *et al.* Mutations in GLUT2, the gene for the liver-type glucose transporter, in patients with Fanconi-Bickel syndrome. *Nature Genet.* **17**, 324–326 (1997).
84. De Vos, A. *et al.* Human and rat  $\beta$  cells differ in glucose transporter but not in glucokinase gene expression. *J. Clin. Invest.* **96**, 2489–2495 (1995).
85. Chang-Chen, K. J., Muller, R. & Bernal-Mizrachi, E.  $\beta$ -Cell failure as a complication of diabetes. *Rev. Endocr. Metab. Disord.* **9**, 329–343 (2008).
86. Leto, D. & Saltiel, A. R. Regulation of glucose transport by insulin: traffic control of GLUT4. *Nature Rev. Mol. Cell Biol.* **13**, 383–396 (2012).
87. Hardie, D. G. AMP-activated protein kinase — an energy sensor that regulates all aspects of cell function. *Genes Dev.* **25**, 1895–1908 (2011).
88. Hardie, D. G., Ross, F. A. & Hawley, S. A. AMPK: a nutrient and energy sensor that maintains energy homeostasis. *Nature Rev. Mol. Cell Biol.* **13**, 251–262 (2012).
89. Efeyan, A. *et al.* Regulation of mTORC1 by the Rag GTPases is necessary for neonatal autophagy and survival. *Nature* **493**, 679–683 (2013).
90. Zoncu, R. *et al.* mTORC1 senses lysosomal amino acids through an inside-out mechanism that requires the vacuolar  $H^+$ -ATPase. *Science* **334**, 678–683 (2011).
91. Zhang, F. *et al.* Molecular mechanism for the umami taste synergism. *Proc. Natl Acad. Sci. USA* **105**, 20930–20934 (2008).
92. Nelson, G. *et al.* Mammalian sweet taste receptors. *Cell* **106**, 381–390 (2001).  
**This paper reports the identification of T1R2–T1R3 as the sweet taste receptor by means of mouse transgenesis and heterologous expression in cultured cells.**
93. Mace, O. J., Affleck, J., Patel, N. & Kellett, G. L. Sweet taste receptors in rat small intestine stimulate glucose absorption through apical GLUT2. *J. Physiol.* **582**, 379–392 (2007).
94. Dyer, J., Salmon, K. S. H., Zibrik, L. & Shirazi-Beechey, S. P. Expression of sweet taste receptors of the T1R family in the intestinal tract and enteroendocrine cells. *Biochem. Soc. Trans.* **33**, 302–305 (2005).
95. Nettleton, J. A. *et al.* Diet soda intake and risk of incident metabolic syndrome and type 2 diabetes in the Multi-Ethnic Study of Atherosclerosis (MESA). *Diabetes Care* **32**, 688–694 (2009).
96. Klionsky, D. J. Autophagy as a regulated pathway of cellular degradation. *Science* **290**, 1717–1721 (2000).
97. Egan, D. F. *et al.* Phosphorylation of ULK1 (hATG1) by AMP-activated protein kinase connects energy sensing to mitophagy. *Science* **331**, 456–461 (2011).
98. Kim, J., Kundu, M., Viollet, B. & Guan, K.-L. AMPK and mTOR regulate autophagy through direct phosphorylation of Ulk1. *Nature Cell Biol.* **13**, 132–141 (2011).
99. Mammucari, C. *et al.* FOXO3 controls autophagy in skeletal muscle *in vivo*. *Cell Metab.* **6**, 458–471 (2007).
100. He, C. & Klionsky, D. J. Regulation mechanisms and signaling pathways of autophagy. *Annu. Rev. Genet.* **43**, 67–93 (2009).
101. Thoreen, C. C. *et al.* An ATP-competitive mammalian target of rapamycin inhibitor reveals rapamycin-resistant functions of mTORC1. *J. Biol. Chem.* **284**, 8023–8032 (2009).
102. Kuma, A. *et al.* The role of autophagy during the early neonatal starvation period. *Nature* **432**, 1032–1036 (2004).  
**This paper demonstrates the essentiality of autophagy as a crucial mechanism to mobilize internal energy stores and to adapt to the interruption of transplacental nutrient supply in neonates.**
103. Komatsu, M. *et al.* Impairment of starvation-induced and constitutive autophagy in Atg7-deficient mice. *J. Cell Biol.* **169**, 425–434 (2005).
104. Yu, L. *et al.* Termination of autophagy and reformation of lysosomes regulated by mTOR. *Nature* **465**, 942–946 (2010).
105. Naito, T., Kuma, A. & Mizushima, N. Differential contribution of insulin and amino acids to the mTORC1-autophagy pathway in the liver and muscle. *J. Biol. Chem.* **288**, 21074–21081 (2013).
106. Kroemer, G., Mariño, G. & Levine, B. Autophagy and the integrated stress response. *Mol. Cell* **40**, 280–293 (2010).
107. Tsukamoto, S. *et al.* Autophagy is essential for preimplantation development of mouse embryos. *Science* **321**, 117–120 (2008).
108. Komatsu, M. *et al.* Loss of autophagy in the central nervous system causes neurodegeneration in mice. *Nature* **441**, 880–884 (2006).
109. Karsli-Uzunbas, G. *et al.* Autophagy is required for glucose homeostasis and lung tumor maintenance. *Cancer Discov.* **4**, 914–927 (2014).
110. Singh, R. *et al.* Autophagy regulates lipid metabolism. *Nature* **458**, 1131–1135 (2009).
111. Mizushima, N. & Komatsu, M. Autophagy: renovation of cells and tissues. *Cell* **147**, 728–741 (2011).
112. de Cabo, R., Carmona-Gutierrez, D., Bernier, M., Hall, M. N. & Madeo, F. The search for antiaging interventions: from elixirs to fasting regimens. *Cell* **157**, 1515–1526 (2014).

**Acknowledgements** D.M.S. is supported by grants from the National Institutes of Health (R01 CA129105, CA103866 and AI047389; R21 AG042876) and awards from the American Federation for Aging, Starr Foundation, Koch Institute Frontier Research Program, and the Ellison Medical Foundation. A.E. is supported by the Charles King's Trust Foundation/Simeon J. Fortin Fellowship. W.C.C. is supported by American Cancer Society – Ellison Foundation Postdoctoral Fellowship (PF-13-356-01-TBE). D.M.S. is an investigator of the Howard Hughes Medical Institute.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at [go.nature.com/kylwoe](http://go.nature.com/kylwoe). Correspondence should be addressed to D.M.S. ([sabatini@wi.mit.edu](mailto:sabatini@wi.mit.edu)).



# Necroptosis and its role in inflammation

Manolis Pasparakis<sup>1</sup> & Peter Vandenabeele<sup>2,3,4</sup>

**Regulated cell death has essential functions in development and in adult tissue homeostasis. Necroptosis is a newly discovered pathway of regulated necrosis that requires the proteins RIPK3 and MLKL and is induced by death receptors, interferons, toll-like receptors, intracellular RNA and DNA sensors, and probably other mediators. RIPK1 has important kinase-dependent and scaffolding functions that inhibit or trigger necroptosis and apoptosis. Mouse-model studies have revealed important functions for necroptosis in inflammation and suggested that it could be implicated in the pathogenesis of many human inflammatory diseases. We discuss the mechanisms regulating necroptosis and its potential role in inflammation and disease.**

Cell death is intricately connected with life in multicellular organisms. The balance between cell death, proliferation and differentiation is crucial for the maintenance of tissue homeostasis throughout life. Programmed cell death (PCD) is essential for many physiological processes, including the shaping of developing organs, epithelial cell renewal and lymphocyte selection. However, cell death that is not developmentally programmed is a sign of stress, injury or infection and is linked to tissue damage and disease pathogenesis. Inflammation is a reaction of the immune system induced in response to infection or tissue injury that is essential for efficient host defence and tissue repair. However, uncontrolled excessive and/or prolonged inflammatory responses cause tissue damage and contribute to the pathogenesis of acute and chronic inflammatory diseases. Although recognized more than 150 years ago as a central component of inflamed tissues, the potential role of cell death as an active component that contributes to tissue homeostasis, inflammation and disease pathogenesis has only recently gained attention<sup>1</sup>. Many of the receptors involved in inflammation can also induce cell death in addition to their potent capacity to drive inflammatory cytokine expression. Recent findings suggest that in addition to the transcriptional regulation of inflammatory genes, the cell-death-inducing properties of these receptors also crucially contribute to inflammation.

For many years apoptosis was considered to be the only form of regulated cell death (RCD), whereas necrosis was seen as an unregulated accidental cell death (ACD) process. Genetic, biochemical and functional evidence, and the discovery of specific chemical inhibitors of necrosis have redefined this process as a molecularly controlled regulated form of cell death<sup>2</sup>. Regulated necrosis includes several cell-death modalities such as necroptosis, parthanatos, ferroptosis or oxytosis, mitochondrial permeability transition (MPT)-dependent necrosis, pyroptosis and pyronecrosis, and cell death associated with the release of (neutrophil) extracellular traps, which is described as NETosis or ETosis (Box 1). Pyroptosis and necroptosis are the best characterized forms of regulated necrosis. Pyroptosis is an inflammatory form of cell death induced by inflammasome activation and has important functions in host defence and inflammation<sup>3</sup>. Necroptosis, mediated by receptor interacting protein kinase-3 (RIPK3) and its substrate mixed lineage kinase like (MLKL), is the best-characterized form of regulated necrosis. Recent studies have provided new and exciting insights into the mechanisms controlling necroptosis and its *in vivo* relevance and suggested that necroptosis could

have important functions in the pathogenesis of several human diseases. This Review discusses the mechanisms controlling regulated necrosis and its physiological relevance for tissue homeostasis and inflammation, focusing particularly on necroptosis.

## Mechanisms and regulation of necroptosis

Much of our knowledge of necroptosis comes from studies of tumour necrosis factor (TNF) signalling. TNF is a pleiotropic cytokine that has a key role in inflammation induced by infection or tissue injury. TNF signalling, primarily through TNF receptor 1 (TNFR1), induces the expression of many genes that regulate inflammation, but under some conditions TNF is also a potent inducer of cell death<sup>4</sup>. Despite its name and early evidence that TNF also induces caspase-independent cell death<sup>5</sup> by a mechanism involving RIPK1 (ref. 6), for many years most studies of TNF-induced cell death focused on apoptosis. The identification of necrostatins as necrosis inhibitors targeting RIPK1 provided evidence that TNF-induced necrosis is a kinase-regulated process, and it was dubbed necroptosis<sup>7,8</sup>. A key step in unravelling the necroptosis pathway was the discovery of RIPK3 as an essential regulator of TNF-induced necrosis<sup>9–11</sup>. The RIP homotypic interaction motif (RHIM) on RIPK3 and RIPK1 allows their interaction and is required for necroptosis induction<sup>9,10</sup>. The necrosome was defined as the complex containing RIPK1 and RIPK3 that was involved in the initiation of necroptosis<sup>12</sup>. RIPK1 and RIPK3 were later shown to form large amyloid-like structures<sup>13</sup>, although it is unclear whether these represent a real signalling platform or a post-event accumulation of these two interacting kinases. The important physiological role of necroptosis was highlighted by a number of genetic studies showing that caspase-8 or Fas-associated protein with death domain (FADD) deficiency cause embryonic lethality and trigger inflammation *in vivo* by sensitizing cells to RIPK3-mediated necroptosis<sup>14–18</sup>.

## Execution of necroptosis

The identification of MLKL pseudokinase as a substrate of RIPK3 required for necroptosis sheds light on the mechanisms involved in executing necrotic cell death downstream of RIPK3 (refs 19, 20). Although initially MLKL was proposed to be associated with the regulation of mitochondrial fission through the proteins phosphoglycerate mutase family member 5 (PGAM5) and dynamin-related protein 1 (DRP1)<sup>21</sup>, the causality of mitochondrial fission in necroptosis has been challenged<sup>22,23</sup>.

<sup>1</sup>Institute for Genetics, Centre for Molecular Medicine and Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases, University of Cologne, 50674 Cologne, Germany.

<sup>2</sup>VIB Inflammation Research Center, Ghent University, UGhent-VIB Research Building FSVM, 9052 Ghent, Belgium <sup>3</sup>Department of Biomedical Molecular Biology, Ghent University, 9000 Ghent, Belgium. <sup>4</sup>Methusalem program, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium.

## BOX 1

## Regulated necrosis pathways

Several types of regulated necrosis exist in addition to necroptosis (reviewed in ref. 2). Parthanatos involves hyperactivation of poly(ADP-ribose) (PAR) polymerase 1 (PARP1), an enzyme originally characterized by its role in DNA-repair mechanisms following a DNA-damage response. The massive PARylation of target proteins leads to cellular depletion of NAD<sup>+</sup> (and consequently of ATP), resulting in a bioenergetic crisis and a form of regulated necrosis. The term ferroptosis was recently coined to describe a type of regulated necrosis that is characterized by iron-dependent production of reactive oxygen species (ROS), which can be blocked by the iron chelator desferrioxamine (DFO). It is elicited by pharmacological inhibition of the antiporter system x<sub>c</sub><sup>-</sup>, which exchanges extracellular cystine for intracellular glutamate. Glutamate toxicity on neurons works by blocking the same antiporter system and has been termed oxytosis. Mitochondrial permeability transition (MPT)-mediated regulated necrosis is another cell-death modality. Cyclophilin D (CYPD) is the sole genetically confirmed component of the permeability transition pore complex (PTPC), which is implicated in MPT-mediated regulated necrosis. Pyroptosis and pyronecrosis are highly inflammatory cell-death modalities characterized by cellular swelling and plasma membrane permeabilization. Pyroptosis occurs after canonical and non-canonical inflammasome stimulation,

leading to caspase-1 and caspase-11 activation, respectively. Some reports distinguish pyronecrosis as a cell-death modality occurring during infection that does not depend on caspase-1 or caspase-11, but requires cathepsin B release following lysosomal membrane permeabilization. NETosis/ETosis (NET is neutrophil extracellular trap; ET is extracellular trap) is also implicated in protection against microbial and viral infection. This modality occurs in neutrophils, eosinophils, mast cells and macrophages. It is associated with chromatin decondensation and release of NETs, which are composed of DNA, chromatin and histones, and allow immune cells to immobilize and kill infectious agents. Besides apoptosis and the various forms of regulated necrosis, a third modality is often put forward: autophagic cell death. The name is slightly misleading because autophagy is in the first instance a crucial mechanism in cellular homeostasis and adaptive responses, however, it can also become a cell-death mechanism. Autophagic cell death is biochemically characterized by markers of autophagy such as lipidation of LC3 (microtubule-associated protein 1 light chain 3) and the degradation of p62, a ubiquitin-binding scaffold protein, and blocked by genetically or pharmacologically targeting the members of the autophagy pathway. Inhibitors that interfere with cell death modalities are shown in Table 1.

**Table 1 | Inhibitors that interfere with cell-death modalities**

Apoptosis		Regulated necrosis					
<b>Morphology</b>							
Cytoplasmic shrinkage		Increasingly translucent cytoplasm					Loss of nuclear integrity Massive chromatin decondensation
Chromatin condensation (pyknosis)		Swelling of organelles; lysosomal membrane permeabilization					
Nuclear fragmentation (karyorrhexis)		Increased cell volume (oncosis)					
Blebbing of the plasma membrane		Permeabilization of the plasma membrane					
Shedding of apoptotic bodies		Mild chromatin condensation; nuclei remain intact					
<b>Death modality</b>							
Intrinsic apoptosis	Extrinsic apoptosis	Necroptosis	Ferroptosis	MPT-mediated regulated necrosis	Parthanatos	Pyroptosis	NETosis/ETosis
<b>Death regulatory factors</b>							
BID, BAX/BAK Cytochrome c APAF1 CASP9	RIPK1* RIPK3† FADD CASP8	RIPK1† RIPK1* RIPK3*	GPX4	CYPD	PARP1	Inflammasome Canonical NLR–ASC–CASP1 Non-canonical (sensor/ adaptor?)-CASP11	NOX
<b>Death execution factors</b>							
CASP3 CASP7		MLKL ion channels	GSH decrease Fe <sup>2+</sup>	Ca <sup>2+</sup> increase	NAD <sup>+</sup> increase ATP increase	ROS increase	
		Lipid peroxidation, energetic catastrophe, and lysosomal and plasma-membrane permeabilization					
<b>Synthetic inhibitor (factor they inhibit)</b>							
zVAD-fmk (CASP) q-VD-Oph (CASP)		NEC1 and NEC1s GSK Cpd27 (RIPK1) GSK843 (RIPK3) GSK872 (RIPK3) NSA (hMLKL) GW906742X (MLKL)	Fer-1 DFO	SfA (CYPD) CsA (CYPD)	3-AB (PARP1) PJ-34 (PARP1)	VX-740 (CASP1) VX-765 (CASP1)	DPI (NOX) GKT137831 (NOX1 and NOX4)
<b>Physiology</b>							
Controlling cell numbers during embryogenesis and homeostasis Immune regulation Pathogen defence		Embryogenesis? Homeostasis? Inflammation IR-injury Thrombosis Neurodegen.	Glu toxicity IR-injury Neurodeg. Transplant.	IR-injury Thrombosis Transplant.	DNA damage Neurodegen.	Inflammation Pathogen defence	Inflammation Extracellular trap formation Pathogen defence

3-AB, 3-aminobenzamide; APAF1, apoptotic protease-activating factor 1; BAK, apoptosis regulator BAK; BAX, apoptosis regulator BAX; BID, BH3-interacting domain death agonist; CASP, caspase; Cpd27, compound 27 or fluoro[2,3-d]pyrimidine 27; CsA, cyclosporine A; DPI, diphenylene iodonium; FADD, Fas-associated death domain; Fer-1, ferrostatin-1; Glu, glutamate; GSH, reduced glutathione; GPX4, glutathione peroxidase 4; MLKL, mixed lineage kinase like; NSA, necrosulfonamide MLKL inhibitor; NEC1, necrostatin-1; NEC1s, necrostatin-1s; Neurodegen., neurodegeneration; NLR, NOD-like receptor; NOX, NADPH oxidase; RIPK, receptor-interacting serine/threonine-protein kinase; qVD-Oph, quinolyl-Val-Asp-Oph; SfA, sanglifehrin A; Transplant., transplantation; TRPM7, transient receptor potential cation channel subfamily M member 7; zVAD-fmk, carbobenzoxy-valyl-alanyl-aspartyl-[O-methyl]-fluoromethylketone. \*Kinase active form, <sup>†</sup>kinase inactive form.

Moreover, depletion of mitochondria by parkin RBR E3 ubiquitin protein ligase (PARK2) and/or carbonyl cyanide *m*-chlorophenylhydrazone (CCCP)-induced mitophagy activation did not prevent TNF-induced necroptosis, arguing against mitochondrial reactive oxygen species (ROS) as a final executioner mechanism<sup>24</sup>. Evidence suggests that RHIM-dependent oligomerization and intramolecular autophosphorylation of RIPK3 results in the recruitment and phosphorylation of MLKL<sup>25,26</sup>, which leads to a conformational change in the pseudokinase domain leading to the exposure of the 4-helical bundle domain<sup>22</sup>. Two non-exclusive models are proposed for the executioner mechanism of MLKL: one as a platform at the plasma membrane for the recruitment of Ca<sup>2+</sup> or Na<sup>+</sup> ion channels<sup>27,28</sup>, and one as a direct pore-forming complex that is recruited through binding of the amino-terminus of the 4-helical bundle domain of MLKL to negatively charged phosphatidylinositol phosphates<sup>29–31</sup>.

### Mechanisms regulating TNFR1-induced necroptosis

In most cells, TNFR1 stimulation is not cytotoxic and induces direct pro-inflammatory signalling through the formation of a membrane-associated protein complex (complex I) (Fig. 1a)<sup>4</sup>. In sensitized cells, TNFR1 induces apoptosis through cytosolic complexes IIa and IIb, and, in particular conditions or cells, necroptosis by the necrosome can be executed. TNF-induced signalling towards necroptosis is prevented by several brakes on RIPK1, as the presence or absence of RIPK1 and its post-translational modifications such as ubiquitylation and phosphorylation are imperative for the biological outcome. RIPK1 comes in several flavours in the four different complexes (I, IIa, IIb and IIc/necrosome), and these are induced in a dynamic way after TNF binding to TNFR1. Ligation of TNFR1 induces the formation of complex I, comprising TNFR1-associated death domain protein (TRADD), RIPK1 and the E3 ubiquitin ligases TNF-receptor-associated factor 2 (TRAF2), the cellular inhibitors of apoptosis (cIAP1 or cIAP2) and the linear ubiquitin chain assembly complex (LUBAC, consisting of haem-oxidized IRP2 ubiquitin ligase-1, HOIL-1L; HOIL-1-interacting protein, HOIP; and SHANK-associated RH domain-interacting protein, SHARPIN). TRADD is required for the recruitment of TRAF2, cIAP1/2 and LUBAC, and for the ubiquitylation of RIPK1 with K63 and linear chains within complex I. This complex favours proinflammatory signalling and prevents cell death through the recruitment of the TGF-activated kinase 1 (TAK1)–TAK1-binding protein (TAB) complex and of the IκB kinase (IKK) complex consisting of IKK1, IKK2 and NF-κB essential modulator (NEMO), leading to NF-κB and mitogen-activated protein kinase (MAPK) activation (reviewed in ref. 32).

Destabilization of complex I leads to the formation of a second cytosolic complex IIa, consisting of TRADD, FADD and caspase-8, which signals towards apoptosis<sup>33–35</sup>. In conditions such as TNF stimulation in the presence of IAP inhibitors (Smac mimetics) or knockdown of IAPs<sup>35</sup>, TAK1 inhibition or knockdown<sup>36</sup>, NEMO knockdown<sup>37</sup> or Pellino knockdown<sup>38</sup>, a cytosolic complex IIb forms that is composed of RIPK1, RIPK3, FADD and caspase-8. This composition resembles the cytosolic ripoptosome complex, which forms independently of TNF, Fas ligand (FASL) or TNF-related apoptosis-inducing ligand (TRAIL) following the loss or inhibition of IAPs<sup>39,40</sup>. This complex IIb favours RIPK1-kinase-activity-dependent apoptosis, however when the levels of RIPK3 and MLKL are sufficiently high and caspase-8 activity is reduced, blocked or absent, complex IIb may evolve to form the necrosome. Caspase-8 is reported to inhibit necroptosis by cleaving RIPK1 (ref. 41) and RIPK3 (ref. 42), as well as CYLD<sup>43</sup>, a deubiquitinating enzyme that removes ubiquitin chains from RIPK1 and contributes to necroptosis *in vitro* and *in vivo*<sup>17,18,44</sup>, but the contribution of the cleavage of each of these proteins to necroptosis inhibition remains unclear. Expression levels of cellular FADD-like interleukin (IL)-1β-converting enzyme (FLICE)-inhibitory protein (FLIP<sub>L</sub>) are crucial in the control of necroptosis and apoptosis. The presence of high levels of FLIP<sub>L</sub> leads to heteromeric caspase-8–FLIP<sub>L</sub> in complex II, which is catalytically active but does not lead to full processing of p10–p20 caspase-8 (and consecutive apoptosis) and alters the substrate specificity<sup>45</sup>.

Heteromeric caspase-8–FLIP<sub>L</sub> prevents complex IIa-dependent apoptosis. The precise mechanism of the paradoxical prosurvival role of caspase-8 in inhibiting necroptosis is not completely clear, but requires the presence of FLIP<sub>L</sub><sup>15</sup>, catalytically active caspase-8 (ref. 46), but not its proteolytic processing<sup>47</sup> and suppression of RIPK1–RIPK3 activation<sup>15</sup>. Thus, FADD–caspase-8–FLIP<sub>L</sub>-mediated control of complex IIb is a second important break preventing the induction of necroptosis. The existence of these brakes (IAPs, TAK1, caspase-8 and FLIP<sub>L</sub>) explains why in most studies of necroptosis they are blocked by a combination of Smac mimetics or TAK1 inhibitor (inhibiting brake 1) and zVAD-fmk (carbobenzoxy-valyl-alanyl-aspartyl-[O-methyl]-fluoromethylketone, inhibiting brake 2), facilitating the efficient formation of the necrosome as long as RIPK3 and MLKL are sufficiently expressed. Most cells are not sensitive to TNF-induced cell death. Adding the protein synthesis inhibitor cycloheximide renders cells sensitive for complex IIa-mediated apoptosis<sup>35</sup>, and, when caspase-8 is inhibited, to necrosome-mediated necroptosis<sup>9</sup>, but the precise protective proteins targeted by cycloheximide remain unclear.

### Other stimuli inducing necroptosis

There are six human death receptors (DRs) in the TNF superfamily: TNFR1, FAS (also known as CD95 or APO-1), DR3 (also known as TRAMP or APO-3), TRAILR1 (also known as DR4), TRAILR2 (also known as DR5, TRICK or KILLER), and DR6 (ref. 48). In contrast to TNFR1 signalling, in which the prosurvival signalling complex forms first and the death-inducing complexes form subsequently in sensitized cells (see earlier), binding of FASL to FAS, or of TRAIL to TRAILR1 or TRAILR2 induces the assembly of a membrane-associated death-inducing signalling complex (DISC) through the adaptor protein FADD, leading to recruitment and activation of caspase-8 and consecutive apoptosis (Fig. 1b). Under particular conditions such as the absence of cIAPs, which favours the recruitment of RIPK1 to Fas<sup>49</sup> and the formation of a cytosolic ripoptosome complex IIb<sup>39</sup>, these ligands also mediate necroptosis when caspase-8 is blocked.

Cellular stress, damage and infection are sensed by several receptors such as Toll-like receptors (TLRs), nucleotide-binding and oligomerization domain (NOD)-like receptors (NLRs), retinoic acid-inducible gene I (RIG-I)-like receptors (RGRs), ripoptosome and protein kinase R (PKR) complexes. Some of these receptors have also been demonstrated to induce necroptosis. Activated TLR3 forms an endosome platform recruiting a cytosolic adaptor, Toll/IL-1 receptor (TIR) domain-containing adaptor protein inducing interferon (IFN)-β (TRIF), which is involved in NF-κB activation and induction of type I IFNs<sup>50</sup>. Besides signalling through the myeloid differentiation primary response 88 (MyD88) adaptor, TLR4 also signals through TRIF<sup>50</sup>. TRIF contains a RHIM-domain, allowing interaction with RIPK1 and RIPK3. Lipopolysaccharide (LPS)–TLR4 or polyinosine–polycytidylic acid (poly(I:C))–TLR3 stimulation in the presence of zVAD-fmk induce TRIF-mediated necroptosis that depends on RIPK3 and MLKL, but may also proceed in the absence of RIPK1 (refs 51, 52). However, inhibition of RIPK1 kinase activity by necrostatin-1 (Nec1)<sup>51,52</sup> or in knock-in macrophages expressing kinase-inactive RIPK1<sup>D138N</sup> (ref. 53) prevented LPS–poly(I:C)–zVAD-fmk-induced necroptosis. These results suggest that RIPK1 is probably recruited to the TRIF-signalling complex and when its kinase activity is inhibited it acts as a blocker of TLR3- or TLR4-induced RIPK3 activation, although TRIF may also directly recruit and activate RIPK3 in the absence of RIPK1 (Fig. 1c)<sup>51,53</sup>. Another RHIM-domain-containing protein is the cytosolic DNA-dependent activator of IFN regulatory factors (DAI; also known as Z-DNA binding protein 1, ZBP1), which in response to viral double-stranded DNA induces NF-κB activation and type I IFNs, but also RIPK3-mediated necroptosis<sup>54</sup> (Fig. 1d).

In the absence of caspase-8 and FADD, both type I and type II IFNs were recently shown to induce RIPK1–RIPK3 necrosome formation in mouse embryonic fibroblasts (MEFs) that depends on the viral RNA responsive protein kinase R, which interacts with RIPK1 (ref. 55). However, independent experiments in PKR-deficient macrophages do not

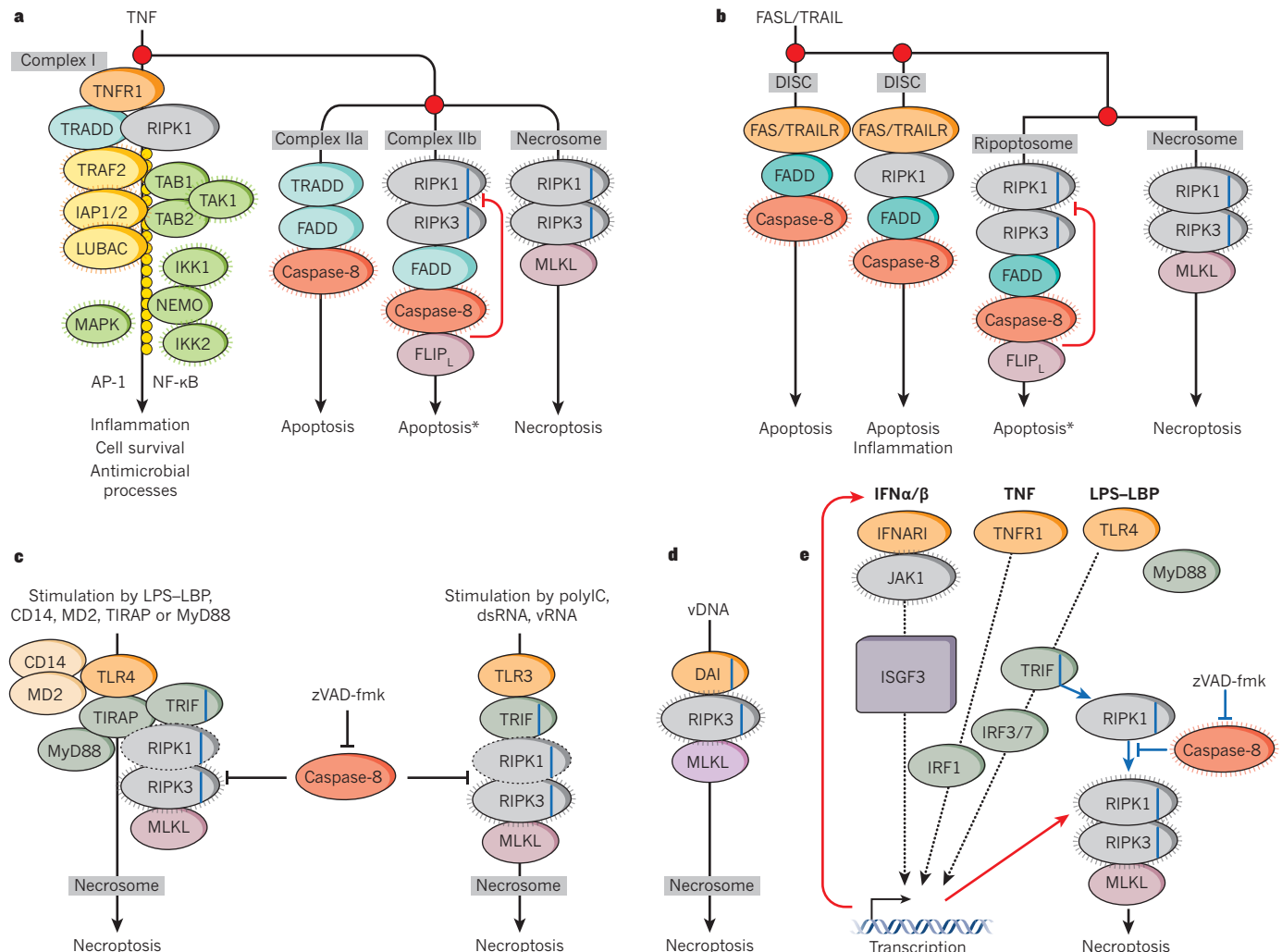


support an essential role for this kinase in IFN- $\alpha$  receptor type I-mediated necroptosis. Interestingly, depending on the cell type, autocrine loops may be implicated in the regulation of necroptosis. In macrophages, LPS-TLR4-, TNF-TNFR1- and poly(I:C)-TLR3-mediated necroptosis requires IFN- $\alpha$  receptor type I signalling, suggesting an autocrine loop of type I IFNs<sup>56</sup> (Fig. 1e). Moreover, stimulation of Pam3CysK-TLR2, Flagellin-TLR5, CpG-TLR9 (ref. 52) or etoposide administration<sup>57</sup> leads to an auto- or paracrine TNF-production loop, which in conditions of IAP inhibition or caspase-8 depletion sensitizes cells to TNF-induced death by RIPK1-dependent necroptosis. Therefore, regulation of necroptosis

could be a crucial factor in controlling responses to infection, DNA damage and inflammation.

### RIPK1 determines cell survival or death

Genetic studies have demonstrated the pivotal role of RIPK1 in cell survival and death. RIPK1 deficiency causes perinatal lethality<sup>34</sup>, which is fully prevented by the combined absence of caspase-8 and RIPK3 (refs 58–60). Caspase-8 deficiency did not prolong survival of *Ripk1*<sup>-/-</sup> mouse neonates, although it prevented apoptosis in many tissues including the intestine, thymus, liver and lung, whereas RIPK3 or MLKL



**Figure 1 | Pathways leading to necroptosis.** **a**, The binding of TNF to its receptor (TNFR1) induces the formation of the receptor-bound complex I (made up of TRADD, RIPK1, TRAF2, IAP1, IAP2 and LUBAC) that activates NF- $\kappa$ B and AP-1 and mitogen-activated protein kinase (MAPK) signalling by the ubiquitin-chain-dependent recruitment of the IKK (made up of NEMO, IKK1 and IKK2) and TAB/TAK-1 complexes. In CHX-treated cells, the cytoplasmic complex IIa (made up of TRADD, FADD and caspase-8) forms, which leads to caspase-8-mediated apoptosis independent of RIPK1. In cells treated with IAP antagonists, TAK1 inhibition or knockdown, or NEMO knockdown the cytoplasmic complex IIb (riposome-like) forms (made up of RIPK1, RIPK3, FADD and caspase-8), resulting in caspase-8-dependent apoptosis, which depends on RIPK1 kinase activity and an RIPK3 platform (apoptosis\*). FLIP<sub>L</sub> keeps caspase-8 in a heteromeric complex that controls RIPK1 and RIPK3 levels by proteolysis. When caspase-8 is inhibited complex IIc or the necrosome (made up of RIPK1, RIPK3 and MLKL) is formed, inducing RIPK1 kinase activity and RIPK3-kinase-activity-dependent necroptosis. **b**, Stimulation of Fas or TRAILR induces the formation of the receptor-bound death-inducing signalling complex (DISC) that triggers caspase-8-mediated apoptosis independent of RIPK1. In the presence of IAP antagonists, Fas and TRAILR stimulation results in the recruitment of RIPK1 generating a complex

similar to complex I in (a), which can progress either in the formation of the cytosolic riposome that induces caspase-8-mediated apoptosis that depends on RIPK1 kinase activity or, when caspase-8 activity is inhibited, RIPK1-kinase-activity-dependent necroptosis. **c**, TLR4 or TLR3 stimulation triggers formation of the necrosome through the RHIM-containing adapter TRIF, resulting in RIPK3-dependent necroptosis in which the role of RIPK1 depends on the cellular context. **d**, DNA-dependent activator of IFN regulatory factors (DAI) recognizes viral double-stranded DNA and through its RHIM domain recruits RIPK3 and induces the formation of the necrosome without RIPK1 and triggers RIPK1-independent RIPK3-kinase-activity-dependent necroptosis. **e**, In bone-marrow-derived macrophages type I interferons (IFN $\alpha$  and  $\beta$ ) induce necroptosis through their cognate receptor, IFNARI, leading to activation of JAK1. These cause the formation of the ISGF3 complex (STAT1–STAT2–IRF9), which in a transcription-dependent way causes induction and activation of the necrosome complex. TNF (IRF1) and lipopolysaccharide (LPS; IRF3/7) signalling can induce an autocrine loop for IFN $\beta$ . In macrophages, IFNARI is required for TNF- and LPS-induced necroptosis. In the presence of zVAD-fmk, sustained activation of the necrosome eventually results in necroptosis. The 'glowing' symbols represent enzymatic activity. Homotypic interaction motif (RHIM domain) is indicated by the blue line.

deficiency ameliorated systemic inflammation, prevented epidermal hyperplasia and marginally prolonged the survival of *Ripk1*<sup>-/-</sup> pups, revealing distinct functions for apoptosis and necroptosis in the multi-organ pathology and perinatal death of *Ripk1*<sup>-/-</sup> mice<sup>58,60</sup>. In addition, intestinal epithelial cell (IEC)-specific knockout of *Ripk1* caused severe lethal intestinal pathology due to FADD–caspase-8-mediated apoptosis of IECs induced primarily, but not exclusively, by TNF<sup>61,62</sup>. Importantly, raising IEC-specific *Ripk1*-knockout mice or *Ripk1*<sup>-/-</sup> *Ripk3*<sup>-/-</sup> mice under germ-free conditions did not prevent apoptosis of IECs and intestinal pathology, demonstrating that the microbiota is not essential to trigger the death of RIPK1-deficient epithelial cells<sup>60,61</sup>, although antibiotic studies suggested that bacteria may aggravate the phenotype<sup>62</sup>. Epidermal-keratinocyte-specific *Ripk1* knockout, however, triggered severe skin inflammation by sensitizing keratinocytes to RIPK3–MLKL-dependent necroptosis, revealing a novel role for RIPK1 as an inhibitor of necroptosis in keratinocytes<sup>61</sup>. Furthermore, RIPK1 deficiency in haematopoietic cells caused bone marrow failure owing to haematopoietic cell apoptosis and necroptosis<sup>60,63</sup>. Collectively, these studies demonstrated the essential role of RIPK1 in preventing apoptosis and necroptosis *in vivo*. Knock-in mice expressing kinase inactive *Ripk1* alleles did not show postnatal lethality or tissue pathology<sup>53,59,61,62,64,65</sup>, demonstrating that RIPK1 mediates cell survival by kinase-independent scaffolding functions. The pro-survival functions of RIPK1 that are important for preventing IEC apoptosis seem to be independent of NF- $\kappa$ B activation<sup>61,62</sup>. Although the exact mechanisms by which RIPK1 prevents cell death remain to be elucidated, RIPK1 deficiency but not lack of its kinase activity resulted in degradation of cIAP1, FLIP<sub>1</sub> and TRAF2 *in vivo* and in response to TNF stimulation *in vitro*, suggesting that kinase-independent RIPK1 scaffolding properties are important to sustain pro-survival signalling platforms<sup>61,62,66</sup>.

RIPK1 kinase activity is not required for activation of NF- $\kappa$ B and MAPK signalling but mediates cell death by either apoptosis or necroptosis. On the one hand, lack of RIPK1 kinase activity partly inhibited but did not prevent RIPK3-mediated necroptosis of FADD-deficient IECs and keratinocytes, demonstrating the existence of RIPK1-kinase-activity-dependent and -independent pathways that induce necroptosis *in vivo*<sup>61</sup>. On the other hand, lack of RIPK1 kinase activity as well as RIPK3 deficiency protected mice from TNF-induced systemic inflammatory response syndrome (SIRS)<sup>53,64,65,67</sup>, demonstrating that kinase-dependent pro-death functions of RIPK1 rather than direct proinflammatory gene induction are crucial for TNF-induced SIRS, as previously proposed<sup>67</sup>.

### Mechanisms determining apoptosis or necroptosis

Despite the advances in unravelling the pathways that regulate necroptosis, the precise mechanisms determining the decision whether a cell will die by apoptosis or necroptosis remain poorly understood. On the positive regulatory side, several studies have suggested that expression levels of RIPK3 and MLKL correlate with sensitivity to necroptosis<sup>10,17,31,60,68–70</sup>. However, a potential drawback of these studies is that RIPK3 and MLKL expression was compared between healthy and inflamed tissues; it is, therefore, difficult to conclude that RIPK3 and/or MLKL overexpression has primary causal functions and is not a secondary consequence of ongoing inflammation. Interestingly, some studies suggest that whether RIPK3 is catalytically active or not may determine necroptosis or apoptosis. As a catalytically inactive platform RIPK3 favours RIPK1-dependent apoptosis, whereas catalytically active RIPK3 induces necroptosis<sup>36</sup>. This bifurcation was illustrated *in vivo* by the finding that, although RIPK3-knockout mice have no spontaneous phenotype, RIPK3<sup>D161N</sup>-kinase-inactive knock-in mice die during embryogenesis due to RIPK1–FADD–caspase-8 mediated apoptosis<sup>65</sup>, suggesting that kinase-active RIPK3 is required to suppress apoptosis. However, a recent report showed that RIPK3<sup>K51A</sup> kinase inactive knock-in mice are viable and that RIPK3 inhibitors block TNF-induced necroptosis, but at higher concentrations induce RIPK3-scaffold-dependent apoptosis, suggesting that RIPK3 conformational changes and not inhibition of its kinase activity trigger apoptosis<sup>71</sup>.

On the negative regulatory side, caspase-8 is the most crucial factor for preventing necroptosis. Indeed, in most *in vivo* experimental systems thus

far, sensitization to necroptosis was achieved by a genetic defect compromising FADD–caspase-8 signalling, and thus inhibiting apoptosis<sup>14–18</sup>. cIAP loss sensitizes cells to apoptosis<sup>72–74</sup> and — when caspase-8 is inhibited — to necroptosis<sup>10</sup>, suggesting that cIAP inhibition confers susceptibility to cell death but that the decision whether a cell dies by apoptosis or necroptosis depends on caspase-8 activity. As already discussed, RIPK1 deficiency sensitizes cells to both apoptosis and necroptosis<sup>58–63</sup>, although the factors determining whether a cell that lacks RIPK1 will die by apoptosis or necroptosis remain elusive. Therefore, more *in vivo* studies are required to understand the mechanisms that determine the susceptibility of different cell types to necroptosis or apoptosis.

### Regulated cell death in inflammation

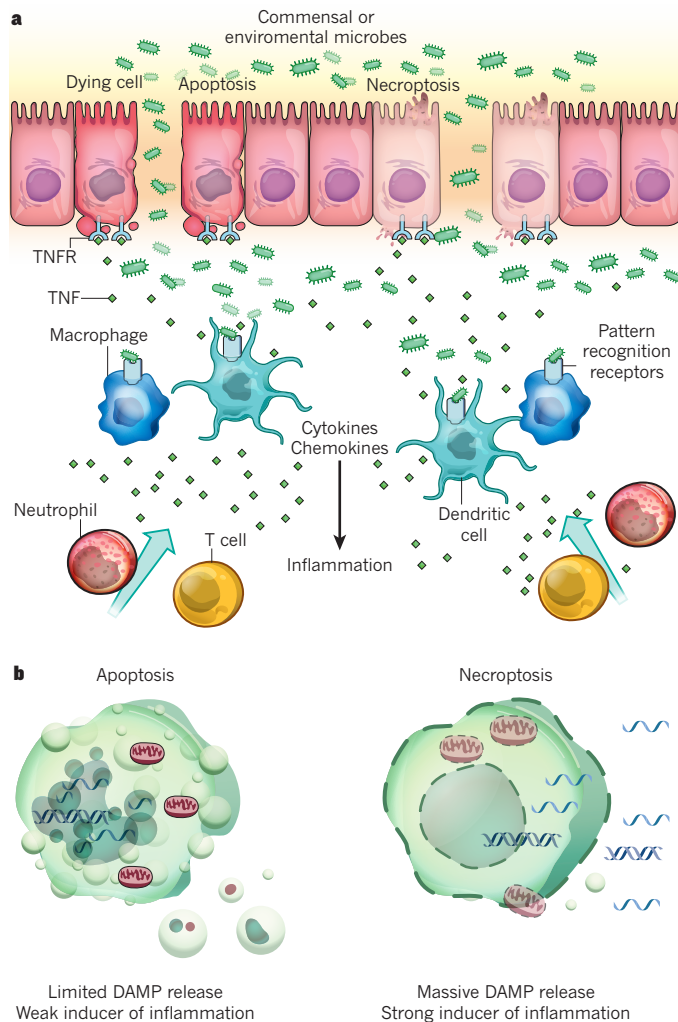
Although initially cell death was considered to be the result of inflammation, more recently the concept that cell death may precede, trigger or amplify the inflammatory response has gained increasing attention<sup>1</sup>. Establishing cell death as an initiator of inflammation *in vivo* is challenging, because it is very difficult to temporally resolve the two responses to demonstrate causality. In the next section, we discuss the evidence implicating necroptosis as an active mediator of inflammation, focusing particularly on the implications of these findings for the pathogenesis of inflammatory diseases — other aspects related to the role of regulated cell death in pathogen defence have been covered in other recent reviews<sup>75,76</sup>.

### Necroptosis as a trigger of inflammation

A major obstacle in studying the role of necroptosis *in vivo* has been the lack of a definite molecular marker for the *in situ* identification of necroptotic cells. Antibodies that detect human phosphorylated MLKL, which were recently reported to detect necrotic hepatocytes in the liver of people with drug-induced liver injury<sup>31</sup>, promise to provide such a tool, but additional validation will be required to demonstrate their applicability for necroptosis detection in other human tissues. The development of mouse-specific phospho-MLKL antibodies will be invaluable for the preclinical validation of phosphorylated MLKL as a marker of necroptotic cells in mouse models of human disease. So far, in most published studies necroptotic cells are described by necrotic morphology and concomitant absence of caspase-3 activation, which cannot distinguish between different types of necrotic cell death (Box 1). In the absence of a specific molecular marker, the only definite criterion for necroptosis has been dependence on RIPK3 and MLKL. Although NEC1 has been used extensively as an inhibitor of necroptosis, in light of the most recent findings that show the existence of RIPK1-kinase-activity-dependent and -independent necroptosis *in vivo*<sup>61</sup>, as well as the complex functions of RIPK1 kinase activity in regulating both necroptosis and apoptosis<sup>77</sup> and the inhibitory effects of NEC1 on indoleamine 2,3-oxygenase<sup>78</sup>, studies based on NEC1 cannot be considered as definite evidence for necroptosis and need to be reproduced in a RIPK3- and/or MLKL-deficient background. Therefore, in this Review, we will only discuss *in vivo* studies that support a proinflammatory role for necroptosis by demonstrating that cell death and inflammation are prevented by RIPK3 and/or MLKL deficiency.

*In vivo* evidence to support a proinflammatory function for necroptosis was initially provided by studies in mice that lack FADD in IECs, which had increased numbers of necrotic IECs and developed spontaneous colitis and ileitis with loss of Paneth cells<sup>17</sup>. RIPK3 deficiency prevented epithelial cell death and inflammation in both the colon and the small intestine of these mice, providing *in vivo* experimental evidence that RIPK3-mediated necroptosis of epithelial cells causes intestinal inflammation<sup>17</sup>. Germ-free conditions and MyD88 or TNF deficiency prevented colitis, but not Paneth-cell loss and ileitis in mice lacking FADD in IECs, suggesting that different mechanisms trigger RIPK3-mediated epithelial cell necroptosis and inflammation in the colon and the small intestine<sup>17</sup>. IEC-specific knockout of caspase-8 caused Paneth cell loss and ileitis<sup>79</sup>, which was subsequently shown to depend on RIPK3, but not on TNFR1 as initially suggested<sup>80</sup>. Interestingly, in contrast to IEC-specific FADD knockouts, IEC-specific caspase-8 knockout mice did not develop colitis, suggesting that FADD and caspase-8 have different functions in the colonic epithelium.

Mice with keratinocyte-specific FADD deficiency developed RIPK3-dependent keratinocyte necroptosis and skin inflammation, suggesting that epithelial cell necroptosis also triggers inflammation in the skin<sup>18</sup>. Keratinocyte-specific caspase-8 knockout mice also developed skin inflammation, which was suggested to be caused by the loss of caspase-8-dependent suppression of RIG-I–IRF3 signalling<sup>81,82</sup>. However, mitochondrial antiviral-signalling protein (MAVS) deficiency did not prevent skin inflammation in mice with inducible ablation of caspase-8, arguing against an important role for the RIG-I–MAVS–IRF3 axis<sup>83</sup>. By contrast, RIPK3 deficiency prevented skin inflammation in these mice<sup>83</sup>, further supporting the idea that RIPK3-dependent keratinocyte necroptosis triggers inflammation in mice lacking caspase-8 or FADD in the epidermis. In addition, epidermis-specific RIPK1 knockout caused keratinocyte



**Figure 2 | Regulated cell death triggers inflammation.** **a**, Epithelial cell death in barrier tissues can cause barrier disruption, allowing commensal or environmental microbes to invade the tissue. Recognition of microbial pathogen-associated molecular patterns (PAMPs) by pattern recognition receptors on myeloid or stromal cells induces the expression of cytokines and chemokines that attract and activate immune cells resulting in inflammation. Cytokines expressed by immune cells (for example, TNF) could trigger the death of additional epithelial cells, further compromising the barrier in a vicious circle, resulting in chronic non-resolving inflammation. In this setting, DAMP release may have a limited contribution as the presence of microbial PAMPs could fully drive the inflammatory response. It is unclear whether apoptosis and necroptosis have a differential capacity to induce barrier disruption. **b**, In sterile tissues, apoptotic cell death is considered a weak inducer of inflammation, as the orderly disassembly of the dying cell allows no, or limited, release of DAMPs. By contrast, the massive release of DAMPs by disintegrating necroptotic cells is believed to be a strong trigger inducing inflammation (see Box 2).

necroptosis and skin inflammation that was prevented by RIPK3 or MLKL deficiency<sup>61</sup>, whereas *Ripk1*<sup>−/−</sup> newborn pups developed RIPK3- and MLKL-dependent skin hyperplasia<sup>61</sup>, providing additional evidence that necroptosis triggers inflammation. TNFR1-deficiency ameliorated, but could not fully prevent inflammation in these mice, suggesting that although TNF is a major driver of cell death and inflammation, TNF-independent pathways also contribute in these models.

In addition to intestinal and skin epithelia, necroptosis of several other cell types was shown to trigger inflammation. RIPK3-dependent necrosis of retinal pigment epithelial cells induced inflammation in a mouse model of double-stranded RNA-induced retinal degeneration<sup>84</sup>. Moreover, RIPK3-deficient mice were protected from TNF or TNF- $\alpha$ -induced SIRS, demonstrating that necroptosis has an important role in TNF-induced shock<sup>53,65,67</sup>. Two studies showed that RIPK3-mediated necroptosis contributes to caecal ligation and puncture-induced sepsis<sup>67,85</sup>, although another study failed to confirm the function of RIPK3 or MLKL in this model<sup>86</sup> of polymicrobial sepsis. In addition, RIPK3- and MLKL-dependent necroptosis was shown to exacerbate tissue injury and inflammation in a mouse model of acute pancreatitis<sup>10,86</sup>, whereas RIPK3 deficiency was protective in a mouse model of Gaucher's disease<sup>87</sup>. Furthermore, RIPK3 deficiency ameliorated ischaemia-reperfusion-induced injury and inflammation and prolonged kidney and heart allograft survival<sup>88–90</sup>, suggesting that necroptosis of graft cells could be a crucial factor in triggering inflammation contributing to transplant rejection.

### How does regulated cell death cause inflammation?

Cell death could induce inflammation indirectly, by disrupting epithelial barriers triggering microbe-driven immune responses (Fig. 2a). This is particularly relevant in the intestinal epithelium, which provides a barrier between luminal bacteria and the mucosal immune system. IEC death usually triggers intestinal inflammation in different mouse models and here it is unclear whether apoptosis and necroptosis differ in their capacity to induce inflammation. Because barrier disruption not only depends on cell death but also on the capacity of the remaining cells to proliferate and keep the barrier sealed, it is important to study whether necroptosis and apoptosis differentially regulate key barrier-related properties of neighbouring epithelial cells such as proliferation, migration and adhesion.

In a sterile setting, dying cells could directly trigger inflammation by releasing factors collectively described as damage-associated molecular patterns (DAMPs) (Box 2). Apoptosis is generally considered to be non-immunogenic because the orderly disassembly of apoptotic cells allows no or limited DAMP release (Fig. 2b). However, there is a general consensus that necroptosis directly triggers inflammation by a massive release of DAMPs from the disintegrating cell<sup>91</sup> (Fig. 2b). Although this notion is clearly supported by evidence that DAMPs are released by necrotic cells and that specific DAMPs are important mediators of inflammation *in vivo*<sup>91</sup>, at present there are no *in vivo* studies that provide direct experimental proof that necroptosis-induced inflammation depends on specific DAMPs. Such experimental proof might be difficult to obtain using genetic models owing to functional redundancy between DAMPs, and the important intracellular functions of several DAMPs. Indeed, surprisingly, myeloid-, hepatocyte- or pancreas-specific knockout of HMGB1, a prototypic DAMP, did not ameliorate but instead exacerbated LPS- or injury-induced damage and inflammation owing to an important role for HMGB1 in maintaining genome homeostasis and cell survival and preventing histone release<sup>92–94</sup>. Antibody neutralization may therefore be more suitable for addressing the functional role of specific DAMPs such as HMGB1, although this approach can not address the cellular origin of DAMPs to formally demonstrate that DAMP release by specific necrotic cells causes inflammation. IL-1 family cytokines (IL-1 $\alpha$ , IL-1 $\beta$ , IL-18, IL-33, IL-36 and IL-37) constitute important DAMPs and their role as mediators of necroptosis-induced inflammation needs to be addressed *in vivo* in relevant genetic models. In the absence of conclusive functional *in vivo* studies, the role of specific DAMPs as mediators of necroptosis-induced inflammation is currently largely based on correlative evidence and nice schemes in reviews, and awaits *in vivo* experimental validation.



## BOX 2

# Damage-associated molecular patterns

Damage-associated molecular patterns (DAMPs) collectively describes the molecules and cellular components that are released or exposed by dying, injured or stressed cells and that act as danger signals to alert the immune system. DAMPs include cytokines and alarmins that are released mainly by dying cells such as the interleukin-1 family cytokines IL-1 $\alpha$  and IL-33, as well as the S100 proteins S100A8, S100A9 and S100A12. In addition, several cellular components that are normally found inside the cell performing important and predominantly non-immunological functions, are released only by dying or damaged cells to act as DAMPs. These include nucleic acids and ribonucleoproteins, histones and HMGB family members, heat-shock proteins,

mitochondrial *N*-formyl peptides and DNA, or even intact mitochondria, F-actin, calreticulin, but also molecules such as monosodium urate and ATP (reviewed in ref. 107). DAMPs are usually detected by pattern recognition receptors that activate immune responses by inducing the expression of cytokines and chemokines. DAMPs are primarily released by cells undergoing necrosis, although apoptotic cells were recently shown to expose or release DAMPs<sup>122</sup>. DAMPs may undergo modifications that inhibit their immunogenicity, for example caspase-dependent proteolysis of IL-33 or oxidation of HMGB1 (refs 123–125). Therefore, the type of cell death may determine the nature, quantity and immunogenicity of the DAMPs released.

## Necroptosis-dependent and -independent functions of RIPK3

Genetic rescue experiments using RIPK3 or MLKL knockouts are routinely used to demonstrate that necroptosis triggers inflammation. However, it remains unclear whether in some cases RIPK3 might be capable of triggering inflammation by exerting necroptosis-independent functions. RIPK3 has been implicated in inflammasome activation and IL-1 $\beta$  release in Smac-mimetic-treated macrophages or in caspase-8-deficient dendritic cells<sup>95,96</sup>, apparently in a cell-death-independent manner, although multiple independent studies failed to show a role for RIPK3 in canonical or non-canonical inflammasome activation<sup>97–99</sup>. RNA viruses, however, were recently reported to activate the NLRP3 inflammasome in a RIPK1- and RIPK3-dependent but MLKL-independent manner, suggesting that in response to viral RNA detection RIPK1 and RIPK3 induce inflammasome activation in a necroptosis-independent manner that seems to involve mitochondrial fission and ROS production<sup>100</sup>. These studies suggest that regulation of the inflammasome by RIPK3 may contribute to inflammation, although available *in vivo* data do not support an important role for the inflammasome in RIPK3-dependent inflammation, because, for example, double IL-1 $\alpha$  and IL-1 $\beta$  deficiency did not prevent inflammation in mice with epidermal keratinocyte-specific knockout of caspase-8 (ref. 81). Furthermore, a recent study reported that RIPK3 regulates LPS-induced NF- $\kappa$ B and caspase-1 activation and inflammatory cytokine production in dendritic cells but not in macrophages, and controls dextran-sodium-sulphate-induced colon inflammation<sup>101</sup>.

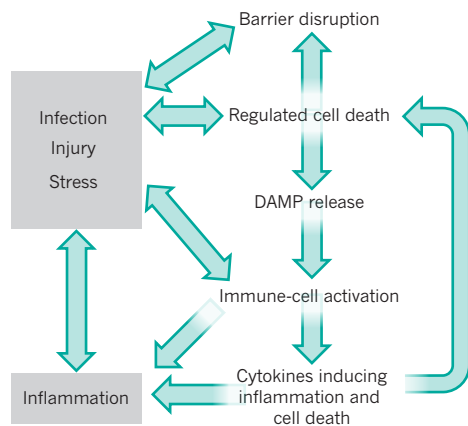
Several studies have shown that caspase-8 activates the inflammasome and also directly cleaves pro-IL-1 $\beta$  to secrete IL-1 $\beta$ <sup>95,97–99,102–104</sup>, suggesting that it could induce inflammation independently of its pro-apoptotic functions. However, most were performed on macrophages lacking both caspase-8 and RIPK3, as caspase-8-knockout macrophages do not survive, raising the question of whether the observed effects might, to some extent, depend on the role of caspase-8 in inhibiting necroptosis. An alternative explanation for these findings could be that inflammasome activation is coupled to cell death in cells that are primed to express inflammasome components and IL-1-family cytokine precursors. Caspase-8 could directly cleave IL-1 precursor proteins and perhaps also caspase-1 during apoptosis, whereas RIPK3–MLKL-dependent necroptosis could trigger inflammasome activation by inducing changes in the redox state, intracellular ion concentrations and the metabolic status of the cell, all of which are well-known inflammasome inducers<sup>3</sup>. This hypothesis is intriguing in light of recent findings showing that IL-1 $\beta$  is released primarily from dying macrophages in response to inflammasome activation<sup>105</sup>. Considering that most stimuli known to act as inflammasome activators in primed macrophages inflict cellular damage, activation of the inflammasome could represent a unifying mechanism endowing primed cells with the capacity to induce inflammation when they are exposed to cell-death-inducing stress. Coupling regulated cell death with inflammasome activation in primed cells could provide a mechanism to ensure that the death of cells exposed to stimuli that activate NF- $\kappa$ B, such as microbial

products or inflammatory cytokines, will trigger an immune response to mediate host defence and promote tissue repair. The potent capacity of most stimuli that trigger regulated cell death (for example, TNF, IFNs and LPS) to also induce proinflammatory gene expression ensures that in a population of cells mixed responses occur between cells that survive or die, and that dying cells have also been primed to produce cytokines and chemokines<sup>106</sup>. Therefore, additional studies in relevant *in vivo* models will be needed to assess the contribution of cell-death-dependent and -independent functions of RIPK3 and caspase-8 in inflammation.

## Is necroptosis more inflammatory than apoptosis?

Apoptosis is generally considered to be non-immunogenic, based on the concept that developmentally programmed cell death, such as that during thymocyte selection in the thymus, should not trigger inflammation. However, apoptotic cell death that is not developmentally programmed indicates tissue injury and should be detected by the immune system to ensure efficient tissue regeneration and host defence. Indeed, several recent studies supported a proinflammatory role of apoptosis. Apoptosis of cancer cells in response to some chemotherapeutic agents was shown to be highly immunogenic, contributing to therapy responses (reviewed in ref. 107). Mouse-model studies suggested that apoptosis of epithelial cells lacking NEMO or TAK1 triggers chronic inflammation<sup>108–111</sup>, although the potential role of necroptosis in these models has not been addressed. Compelling evidence that epithelial cell apoptosis induces inflammation comes from studies showing that inducible FLIP<sub>1</sub> knockout caused keratinocyte apoptosis and inflammation<sup>83,112</sup>, which could not be prevented by RIPK3 deficiency<sup>83</sup>. In addition, skin inflammation in chronic proliferative dermatitis mice was recently shown to depend primarily on TNFR1–TRADD–RIPK1-dependent death of SHARPIN-deficient keratinocytes mediated primarily by FADD–caspase-8-dependent apoptosis, whereas RIPK3-dependent necroptosis contributes to disease severity, but is not essential for the induction of inflammation<sup>64,113,114</sup>.

Necroptosis is believed to be a more potent inducer of inflammation than apoptosis but this concept has not been rigorously tested in relevant *in vivo* experimental models. Clearance of dying cells is crucial in limiting inflammation<sup>115</sup>. Although both apoptotic and necroptotic cells are recognised and cleared by phagocytes, the mechanism may be different owing to profound differences in the morphology, blebbing and apoptotic body formation in apoptosis compared with oncosis and rapid plasma membrane permeabilization in the case of necroptosis. The more acute release of DAMPs could support increased proinflammatory properties of necroptosis; however, as already discussed, the role of specific DAMPs in necroptosis-induced inflammation *in vivo* has not been functionally validated. Particularly considering that co-incubating apoptotic or necrotic cells with macrophages did not result in cytokine induction<sup>116</sup>, the *in vivo* context may be very relevant for this type of experiment. To allow a direct comparison of apoptosis and necroptosis, ideally they would need to be induced in a similar number of cells in the same tissue. Recent



**Figure 3 | Regulated cell death fuels the vicious circle in chronic inflammation.** Regulated cell death could play a part in the initiation and amplification or chronicity of inflammation. This form of cell death can induce immune activation and cytokine expression indirectly through barrier disruption, allowing microbial entry, or directly by the release of DAMPs. Activated immune cells produce cytokines that induce inflammation directly by activating the expression of proinflammatory genes but also trigger regulated cell death, closing a circle that amplifies the inflammatory response. Infection, injury and stress can initiate the response by inducing regulated cell death directly and/or by triggering immune-cell activation and cytokine production.

results showing that necroptosis, but not apoptosis, of RIPK1-deficient keratinocytes triggers skin inflammation provide initial *in vivo* experimental evidence that, at least in this model, keratinocyte necroptosis is a more potent trigger of skin inflammation compared with apoptosis<sup>61</sup>, although, as already discussed, keratinocyte apoptosis has been shown to trigger inflammation in other models. In conclusion, although there are indications that necroptosis is more inflammatory than apoptosis, new and more specific *in vivo* models are needed to directly compare the capacity of these two pathways of regulated cell death to trigger inflammation and address the underlying mechanisms. Such models could take advantage of intracellular mediators that exclusively induce apoptosis (for example, tBid) or necroptosis (for example, constitutively active MLKL mutants) to specifically address the function of each cell-death pathway without interference from extracellular inducers.

### Necroptosis in human disease

Studies in mice have provided strong evidence suggesting that necroptosis has a crucial role in disease pathogenesis. In addition to acting as an initiation signal, regulated cell death could contribute to the amplification and chronicity of inflammation since many cytokines produced during the immune response (such as, TNF family members or INFs) are potent cell-death inducers. A vicious circle of cell death, DAMP release, immune-cell activation and release of death-inducing cytokines may fuel prolonged non-resolving inflammatory responses and contribute to the pathogenesis of chronic inflammatory diseases (Fig. 3). It is intriguing to speculate that the pathogenic role of TNF in chronic inflammation could be, at least in part, due to its cell-death inducing properties. However, few studies addressing the role of necroptosis in human inflammatory diseases exist and these rely largely on correlative evidence such as the upregulation of RIPK3, which is only indicative and does not provide proof that necroptosis takes place and is causally associated with disease pathogenesis.

Epithelial cell death has been associated with intestinal inflammation in inflammatory bowel disease (IBD) and although early reports focused on apoptosis it is not clear whether necroptosis also occurs<sup>117</sup>, in particular because most studies used TdT-mediated dUTP nick end labelling (TUNEL) assays that cannot distinguish between apoptotic and necroptotic cells. Mouse-model studies showed that epithelial-cell necroptosis induces intestinal inflammation, suggesting that necroptosis could also contribute to the pathogenesis of IBD in humans<sup>17,79,83</sup>. However, evidence that necroptosis is implicated in human IBD remains limited.

An initial study identified dying cells with necrotic morphology in ileal crypt sections from patients with Crohn's disease, indicating that Paneth cells undergo necroptosis<sup>79</sup>. A more recent study showed that RIPK3 and MLKL were upregulated, whereas caspase-8 expression was reduced in the inflamed mucosa of children with Crohn's disease, ulcerative colitis or allergic colitis and suggested that necroptosis is a crucial event that amplifies inflammation and contributes to these intestinal pathologies<sup>118</sup>. Although these studies do indicate that necroptosis might be involved in IBD they are based on rather weak and correlative evidence, therefore more research using specific and sensitive molecular markers of necroptosis will be required to establish the potential role of necroptosis in IBD.

Mouse-model experiments identified keratinocyte necroptosis as a potent trigger of skin inflammation<sup>18,61,83</sup>, suggesting that keratinocyte necroptosis might also be implicated in the pathogenesis of human inflammatory skin diseases. A recent report suggested that keratinocyte necroptosis contributes to severe cutaneous adverse drug reactions in humans and in a mouse model of the disease<sup>119</sup>. However, the occurrence and potential role of necroptosis in human inflammatory skin diseases such as psoriasis has not been investigated. Recent reports based on mouse models suggested that RIPK3-mediated necroptosis contributes to liver injury and inflammation induced by alcohol but also in non-alcoholic steatohepatitis<sup>120,121</sup>. Upregulation of RIPK3 was detected in hepatocytes in the livers of people with alcohol-induced liver disease and non-alcoholic steatohepatitis, suggesting that RIPK3-dependent necroptosis could contribute to hepatocyte death and inflammation<sup>120,121</sup>. Moreover, immunostaining with antibodies that recognize phosphorylated MLKL, which promise to provide a specific marker of necroptosis, suggested that hepatocyte necroptosis occurs in the livers of people with drug-induced liver injury<sup>31</sup>. Therefore, although mouse-model experiments strongly suggest that necroptosis could be implicated in inflammatory diseases, more studies using definite molecular markers of necroptosis and functional validation using specific inhibitors will be required to establish the involvement of necroptosis in the pathogenesis of human diseases.

### Outlook

Necroptosis is now established as an important pathway of regulated cell death, but many questions remain to be addressed. What are the mechanisms controlling the formation and activity of the necrosome? What is the role of specific ubiquitylation events and enzymes? What is the precise mechanism of MLKL-mediated necroptosis? How is the kinase activity of RIPK1 and RIPK3 regulated and what are their substrates? What are the specific kinase-dependent and scaffolding functions of RIPK1 controlling inflammation, cell survival, apoptosis and necroptosis? What is the relative contribution of cell-death-dependent and -independent functions of RIPK3 and RIPK1 in inflammation? How does necroptosis contribute to the initiation, amplification and chronicity of inflammation? Does necroptosis display different properties in regulating inflammation compared with other types of regulated cell death? What is the relationship between RIPKs, necroptosis and inflammasome activation? These questions become particularly important when considering that stimuli triggering cell death such as TNF also potentially induce inflammatory gene expression. The clinical efficacy of anti-TNF therapy has established TNF as a key player in chronic inflammatory diseases including IBD, rheumatoid arthritis and psoriasis, but the TNF-dependent pathogenic mechanisms remain unclear. Determining the relative contribution of cell-death-dependent and -independent pathways in TNF-induced chronic inflammation may lead to new and more specific therapeutic targets. New preclinical mouse-model studies will be needed to answer these mechanistic questions. Addressing the role of necroptosis in human disease remains a major challenge and will require the establishment of specific, sensitive and reliable molecular markers of necroptosis. Current data raise the hope that manipulating necroptosis could provide new and urgently needed therapeutic opportunities in acute and chronic inflammatory conditions. It remains to be seen if this promise will be fulfilled. ■

Received 25 September; accepted 11 November 2014.

1. Wallach, D., Kang, T. B. & Kovalenko, A. Concepts of tissue injury and cell death in inflammation: a historical perspective. *Nature Rev. Immunol.* **14**, 51–59 (2014).
2. Vanden Berghe, T., Linkermann, A., Jouan-Lanhouet, S., Walczak, H. & Vandenabeele, P. Regulated necrosis: the expanding network of non-apoptotic cell death pathways. *Nature Rev. Mol. Cell Biol.* **15**, 135–147 (2014).
3. Lamkanfi, M. & Dixit, V. M. Mechanisms and functions of inflammasomes. *Cell* **157**, 1013–1022 (2014).
4. Vandenabeele, P., Galluzzi, L., Vanden Berghe, T. & Kroemer, G. Molecular mechanisms of necroptosis: an ordered cellular explosion. *Nature Rev. Mol. Cell Biol.* **11**, 700–714 (2010).
5. Vercammen, D. *et al.* Inhibition of caspases increases the sensitivity of L929 cells to necrosis mediated by tumor necrosis factor. *J. Exp. Med.* **187**, 1477–1485 (1998).
6. Holler, N. *et al.* Fas triggers an alternative, caspase-8-independent cell death pathway using the kinase RIP as effector molecule. *Nature Immunol.* **1**, 489–495 (2000).
7. Degterev, A. *et al.* Chemical inhibitor of nonapoptotic cell death with therapeutic potential for ischemic brain injury. *Nature Chem. Biol.* **1**, 112–119 (2005).
8. Degterev, A. *et al.* Identification of RIP1 kinase as a specific cellular target of necrostatins. *Nature Chem. Biol.* **4**, 313–321 (2008).
9. Cho, Y. S. *et al.* Phosphorylation-driven assembly of the RIP1–RIP3 complex regulates programmed necrosis and virus-induced inflammation. *Cell* **137**, 1112–1123 (2009).
10. He, S. *et al.* Receptor interacting protein kinase-3 determines cellular necrotic response to TNF- $\alpha$ . *Cell* **137**, 1100–1111 (2009).
11. Zhang, D. W. *et al.* RIP3, an energy metabolism regulator that switches TNF-induced cell death from apoptosis to necrosis. *Science* **325**, 332–336 (2009). **This paper along with refs 9 and 10 identified RIPK3 as a key regulator of TNF-induced necroptosis.**
12. Vandenabeele, P., Declercq, W., Van Herreweghe, F. & Vanden Berghe, T. The role of the kinases RIP1 and RIP3 in TNF-induced necrosis. *Sci. Signal.* **3**, re4 (2010).
13. Li, J. *et al.* The RIP1/RIP3 necrosome forms a functional amyloid signaling complex required for programmed necrosis. *Cell* **150**, 339–350 (2012).
14. Kaiser, W. J. *et al.* RIP3 mediates the embryonic lethality of caspase-8-deficient mice. *Nature* **471**, 368–372 (2011).
15. Oberst, A. *et al.* Catalytic activity of the caspase-8-FLIP<sub>L</sub> complex inhibits RIPK3-dependent necrosis. *Nature* **471**, 363–367 (2011).
16. Zhang, H. *et al.* Functional complementation between FADD and RIP1 in embryos and lymphocytes. *Nature* **471**, 373–376 (2011). **This paper along with refs 14 and 15 showed that RIPK3 knockout prevents embryonic lethality of caspase-8 and FADD, revealing the important role of RIPK3 in inhibiting RIPK3 during development.**
17. Welz, P. S. *et al.* FADD prevents RIP3-mediated epithelial cell necrosis and chronic intestinal inflammation. *Nature* **477**, 330–334 (2011).
18. Bonnet, M. C. *et al.* The adaptor protein FADD protects epidermal keratinocytes from necroptosis *in vivo* and prevents skin inflammation. *Immunity* **35**, 572–582 (2011). **This paper along with ref. 17 showed for the first time that RIPK3-dependent epithelial cell necroptosis triggers inflammation in the intestine and the skin.**
19. Sun, L. *et al.* Mixed lineage kinase domain-like protein mediates necrosis signaling downstream of RIP3 kinase. *Cell* **148**, 213–227 (2012).
20. Zhao, J. *et al.* Mixed lineage kinase domain-like is a key receptor interacting protein 3 downstream component of TNF-induced necrosis. *Proc. Natl Acad. Sci. USA* **109**, 5322–5327 (2012). **This paper along with ref. 19 identified MLKL as an essential mediator of necroptosis downstream of RIPK3.**
21. Wang, Z., Jiang, H., Chen, S., Du, F. & Wang, X. The mitochondrial phosphatase PGAM5 functions at the convergence point of multiple necrotic death pathways. *Cell* **148**, 228–243 (2012).
22. Murphy, J. M. *et al.* The pseudokinase MLKL mediates necroptosis via a molecular switch mechanism. *Immunity* **39**, 443–453 (2013).
23. Remijsen, Q. *et al.* Depletion of RIPK3 or MLKL blocks TNF-driven necroptosis and switches towards a delayed RIPK1 kinase-dependent apoptosis. *Cell Death Dis.* **5**, e1004 (2014).
24. Tait, S. W. *et al.* Widespread mitochondrial depletion via mitophagy does not compromise necroptosis. *Cell Rep.* **5**, 878–885 (2013).
25. Orozco, S. *et al.* RIPK1 both positively and negatively regulates RIPK3 oligomerization and necroptosis. *Cell Death Differ.* **21**, 1511–1521 (2014).
26. Wu, X. N. *et al.* Distinct roles of RIP1–RIP3 hetero- and RIP3–RIP3 homo-interaction in mediating necroptosis. *Cell Death Differ.* **21**, 1709–1720 (2014).
27. Cai, Z. *et al.* Plasma membrane translocation of trimerized MLKL protein is required for TNF-induced necroptosis. *Nature Cell Biol.* **16**, 55–65 (2014).
28. Chen, X. *et al.* Translocation of mixed lineage kinase domain-like protein to plasma membrane leads to necrotic cell death. *Cell Res.* **24**, 105–121 (2014).
29. Su, L. *et al.* A plug release mechanism for membrane permeation by MLKL. *Structure* **22**, 1489–1500 (2014).
30. Dondelinger, Y. *et al.* MLKL compromises plasma membrane integrity by binding to phosphatidylinositol phosphates. *Cell Rep.* **7**, 971–981 (2014).
31. Wang, H. *et al.* Mixed lineage kinase domain-like protein causes necrotic membrane disruption upon phosphorylation by RIP3. *Mol. Cell* **54**, 133–146 (2014).
32. Silke, J. & Brink, R. Regulation of TNFRSF and innate immune signalling complexes by TRAFs and cIAPs. *Cell Death Differ.* **17**, 35–45 (2010).
33. Micheau, O. & Tschopp, J. Induction of TNF receptor I-mediated apoptosis via two sequential signaling complexes. *Cell* **114**, 181–190 (2003). **This study showed that TNFR1-induced inflammatory and apoptotic signalling is mediated by two different signalling complexes that form sequentially at the cell membrane and the cytoplasm.**
34. Kelliher, M. A. *et al.* The death domain kinase RIP mediates the TNF-induced NF- $\kappa$ B signal. *Immunity* **8**, 297–303 (1998).
35. Wang, L., Du, F. & Wang, X. TNF- $\alpha$  induces two distinct caspase-8 activation pathways. *Cell* **133**, 693–703 (2008).
36. Dondelinger, Y. *et al.* RIPK3 contributes to TNFR1-mediated RIPK1 kinase-dependent apoptosis in conditions of cIAP1/2 depletion or TAK1 kinase inhibition. *Cell Death Differ.* **20**, 1381–1392 (2013).
37. Legarda-Addison, D., Hase, H., O'Donnell, M. A. & Ting, A. T. NEMO/I $\kappa$ B $\kappa$  regulates an early NF- $\kappa$ B-independent cell-death checkpoint during TNF signaling. *Cell Death Differ.* **16**, 1279–1288 (2009).
38. Yang, S. *et al.* Pellino3 targets RIP1 and regulates the pro-apoptotic effects of TNF- $\alpha$ . *Nature Commun.* **4**, 2583 (2013).
39. Feoktistova, M. *et al.* cIAPs block Ripoptosome formation, a RIP1/caspase-8 containing intracellular cell death complex differentially regulated by cFLIP isoforms. *Mol. Cell* **43**, 449–463 (2011).
40. Tenev, T. *et al.* The Ripoptosome, a signaling platform that assembles in response to genotoxic stress and loss of IAPs. *Mol. Cell* **43**, 432–448 (2011).
41. Lin, Y., Devin, A., Rodriguez, Y. & Liu, Z. G. Cleavage of the death domain kinase RIP by caspase-8 prompts TNF-induced apoptosis. *Genes Dev.* **13**, 2514–2526 (1999).
42. Feng, S. *et al.* Cleavage of RIP3 inactivates its caspase-independent apoptosis pathway by removal of kinase domain. *Cell. Signal.* **19**, 2056–2067 (2007).
43. O'Donnell, M. A. *et al.* Caspase 8 inhibits programmed necrosis by processing CYLD. *Nature Cell Biol.* **13**, 1437–1442 (2011).
44. Hitomi, J. *et al.* Identification of a molecular signaling network that regulates a cellular necrotic cell death pathway. *Cell* **135**, 1311–1323 (2008).
45. Pop, C. *et al.* FLIP<sub>L</sub> induces caspase 8 activity in the absence of interdomain caspase 8 cleavage and alters substrate specificity. *Biochem. J.* **433**, 447–457 (2011).
46. Lu, J. V. *et al.* Complementary roles of Fas-associated death domain (FADD) and receptor interacting protein kinase-3 (RIPK3) in T-cell homeostasis and antiviral immunity. *Proc. Natl Acad. Sci. USA* **108**, 15312–15317 (2011).
47. Kang, T. B. *et al.* Mutation of a self-processing site in caspase-8 compromises its apoptotic but not its nonapoptotic functions in bacterial artificial chromosome-transgenic mice. *J. Immunol.* **181**, 2522–2532 (2008).
48. Wilson, N. S., Dixit, V. & Ashkenazi, A. Death receptor signal transducers: nodes of coordination in immune signaling networks. *Nature Immunol.* **10**, 348–355 (2009).
49. Geserick, P. *et al.* Cellular IAPs inhibit a cryptic CD95-induced cell death by limiting RIP1 kinase recruitment. *J. Cell Biol.* **187**, 1037–1054 (2009).
50. Takeuchi, O. & Akira, S. Pattern recognition receptors and inflammation. *Cell* **140**, 805–820 (2010).
51. He, S., Liang, Y., Shao, F. & Wang, X. Toll-like receptors activate programmed necrosis in macrophages through a receptor-interacting kinase-3-mediated pathway. *Proc. Natl Acad. Sci. USA* **108**, 20054–20059 (2011).
52. Kaiser, W. J. *et al.* Toll-like receptor 3-mediated necrosis via TRIF, RIP3, and MLKL. *J. Biol. Chem.* **288**, 31268–31279 (2013).
53. Polykratis, A. *et al.* Cutting edge: RIPK1 kinase inactive mice are viable and protected from TNF-induced necroptosis *in vivo*. *J. Immunol.* **193**, 1539–1543 (2014).
54. Upton, J. W., Kaiser, W. J., Mocarski, E. S. DAI/ZBP1/DLM-1 complexes with RIP3 to mediate virus-induced programmed necrosis that is targeted by murine cytomegalovirus vIRA. *Cell Host Microbe* **11**, 290–297 (2012).
55. Thapa, R. J. *et al.* Interferon-induced RIP1/RIP3-mediated necrosis requires PKR and is licensed by FADD and caspases. *Proc. Natl Acad. Sci. USA* **110**, E3109–E3118 (2013).
56. McComb, S. *et al.* Type-I interferon signaling through ISGF3 complex is required for sustained Rip3 activation and necroptosis in macrophages. *Proc. Natl Acad. Sci. USA* **111**, E3206–E3213 (2014).
57. Biton, S. & Ashkenazi, A. NEMO and RIP1 control cell fate in response to extensive DNA damage via TNF- $\alpha$  feedforward signaling. *Cell* **145**, 92–103 (2011).
58. Dillon, C. P. *et al.* RIPK1 blocks early postnatal lethality mediated by caspase-8 and RIPK3. *Cell* **157**, 1189–1202 (2014).
59. Kaiser, W. J. *et al.* RIP1 suppresses innate immune necrotic as well as apoptotic cell death during mammalian parturition. *Proc. Natl Acad. Sci. USA* **111**, 7753–7758 (2014). **This paper, ref. 58 and ref. 59 showed that early postnatal lethality of RIPK1-deficient mice is rescued by double knockout of caspase-8 and RIPK3 and that caspase-8-dependent apoptosis and RIPK3–MLKL-mediated necroptosis differentially contribute to tissue pathologies in RIPK1-deficient pups.**
61. Dannappel, M. *et al.* RIPK1 maintains epithelial homeostasis by inhibiting apoptosis and necroptosis. *Nature* **513**, 90–94 (2014). **This study showed that the kinase-independent function of RIPK1 is essential for intestinal and skin homeostasis by preventing epithelial cell apoptosis and necroptosis.**
62. Takahashi, N. *et al.* RIPK1 ensures intestinal homeostasis by protecting the epithelium against apoptosis. *Nature* **513**, 95–99 (2014). **This study showed that the kinase-independent function of RIPK1 prevents IEC apoptosis.**
63. Roderick, J. E. *et al.* Hematopoietic RIPK1 deficiency results in bone marrow failure caused by apoptosis and RIPK3-mediated necroptosis. *Proc. Natl Acad. Sci. USA* **111**, 14436–14441 (2014).
64. Berger, S. B. *et al.* Cutting edge: RIP1 kinase activity is dispensable for normal development but is a key regulator of inflammation in SHARPIN-deficient mice. *J. Immunol.* **192**, 5476–5480 (2014). **This article along with ref. 53 (and ref. 65) showed that knock-in mice expressing kinase inactive *Ripk1* alleles are viable and protected from TNF-induced SIRS.**
65. Newton, K. *et al.* Activity of protein kinase RIPK3 determines whether cells die by necroptosis or apoptosis. *Science* **343**, 1357–1360 (2014). **This study showed that RIPK3<sup>D161N</sup> knock-in mice die during embryogenesis due to caspase-8 dependent apoptosis, suggesting that RIPK3-kinase activity**



**inhibits caspase-8 during development.**

66. Gentile, I. E. *et al.* In TNF-stimulated cells, RIPK1 promotes cell survival by stabilizing TRAF2 and cIAP1, which limits induction of non-canonical NF- $\kappa$ B and activation of caspase-8. *J. Biol. Chem.* **286**, 13282–13291 (2011).
67. Duprez, L. *et al.* RIP kinase-dependent necrosis drives lethal systemic inflammatory response syndrome. *Immunity* **35**, 908–918 (2011).  
**This study showed for the first time that TNF-mediated SIRS is RIPK3-dependent and blocked by NEC1.**
68. Lin, J. *et al.* A role of RIP3-mediated macrophage necrosis in atherosclerosis development. *Cell Rep.* **3**, 200–210 (2013).
69. Sato, K. *et al.* Receptor interacting protein kinase-mediated necrosis contributes to cone and rod photoreceptor degeneration in the retina lacking interphotoreceptor retinoid-binding protein. *J. Neurosci.* **33**, 17458–17468 (2013).
70. Colbert, L. E. *et al.* Pronecrotic mixed lineage kinase domain-like protein expression is a prognostic biomarker in patients with early-stage resected pancreatic adenocarcinoma. *Cancer* **119**, 3148–3155 (2013).
71. Mandal, P. *et al.* RIP3 induces apoptosis independent of pro-necrotic kinase activity. *Mol. Cell* **56**, 481–495 (2014).  
**This study showed that RIPK3<sup>NEC1</sup> knock-in mice are viable and that RIPK3 kinase inhibitors block necroptosis but at high concentrations induce apoptosis, suggesting that RIPK3 conformational changes and not lack of its kinase activity trigger apoptosis.**
72. Petersen, S. L. *et al.* Autocrine TNF $\alpha$  signaling renders human cancer cells susceptible to Smac-mimetic-induced apoptosis. *Cancer Cell* **12**, 445–456 (2007).
73. Varfolomeev, E. *et al.* IAP antagonists induce autoubiquitination of c-IAPs, NF- $\kappa$ B activation, and TNF $\alpha$ -dependent apoptosis. *Cell* **131**, 669–681 (2007).
74. Vince, J. E. *et al.* IAP antagonists target cIAP1 to induce TNF $\alpha$ -dependent apoptosis. *Cell* **131**, 682–693 (2007).
75. Blander, J. M. A long-awaited merger of the pathways mediating host defence and programmed cell death. *Nature Rev. Immunol.* **14**, 601–618 (2014).
76. Mocarski, E. S., Upton, J. W. & Kaiser, W. J. Viral infection and the evolution of caspase 8-regulated apoptotic and necrotic death pathways. *Nature Rev. Immunol.* **12**, 79–88 (2012).
77. Christofferson, D. E., Li, Y. & Yuan, J. Control of life-or-death decisions by RIP1 kinase. *Annu. Rev. Physiol.* **76**, 129–150 (2014).
78. Takahashi, N. *et al.* Necrostatin-1 analogues: critical issues on the specificity, activity and *in vivo* use in experimental disease models. *Cell Death Dis.* **3**, e437 (2012).  
**This study revealed crucial issues on the specificity and *in vivo* use of NEC1.**
79. Günther, C. *et al.* Caspase-8 regulates TNF- $\alpha$ -induced epithelial necroptosis and terminal ileitis. *Nature* **477**, 335–339 (2011).
80. Wittkopf, N. *et al.* Cellular FLICE-like inhibitory protein secures intestinal epithelial cell survival and immune homeostasis by regulating caspase-8. *Gastroenterology* **145**, 1369–1379 (2013).
81. Kovalenko, A. *et al.* Caspase-8 deficiency in epidermal keratinocytes triggers an inflammatory skin disease. *J. Exp. Med.* **206**, 2161–2177 (2009).
82. Rajput, A. *et al.* RIG-I RNA helicase activation of IRF3 transcription factor is negatively regulated by caspase-8-mediated cleavage of the RIP1 protein. *Immunity* **34**, 340–351 (2011).
83. Weinlich, R. *et al.* Protective roles for caspase-8 and cFLIP in adult homeostasis. *Cell Rep.* **5**, 340–348 (2013).
84. Murakami, Y. *et al.* Programmed necrosis, not apoptosis, is a key mediator of cell loss and DAMP-mediated inflammation in dsRNA-induced retinal degeneration. *Cell Death Differ.* **21**, 270–277 (2014).
85. Sharma, A., Matsuo, S., Yang, W. L., Wang, Z. & Wang, P. Receptor-interacting protein kinase 3 deficiency inhibits immune cell infiltration and attenuates organ injury in sepsis. *Crit. Care* **18**, R142 (2014).
86. Wu, J. *et al.* Mkl1 knockout mice demonstrate the indispensable role of Mkl1 in necroptosis. *Cell Res.* **23**, 994–1006 (2013).
87. Vitner, E. B. *et al.* RIPK3 as a potential therapeutic target for Gaucher's disease. *Nature Med.* **20**, 204–208 (2014).
88. Linkermann, A. *et al.* Two independent pathways of regulated necrosis mediate ischemia-reperfusion injury. *Proc. Natl Acad. Sci. USA* **110**, 12024–12029 (2013).
89. Lau, A. *et al.* RIPK3-mediated necroptosis promotes donor kidney inflammatory injury and reduces allograft survival. *Am. J. Transplant.* **13**, 2805–2818 (2013).
90. Pavlosky, A. *et al.* RIPK3-mediated necroptosis regulates cardiac allograft rejection. *Am. J. Transplant.* **14**, 1778–1790 (2014).
91. Kaczmarek, A., Vandenabeele, P. & Krysko, D. V. Necroptosis: the release of damage-associated molecular patterns and its physiological relevance. *Immunity* **38**, 209–223 (2013).
92. Kang, R. *et al.* Intracellular Hmgb1 inhibits inflammatory nucleosome release and limits acute pancreatitis in mice. *Gastroenterology* **146**, 1097–1107 (2014).
93. Huang, H. *et al.* Hepatocyte-specific high-mobility group box 1 deletion worsens the injury in liver ischemia/reperfusion: a role for intracellular high-mobility group box 1 in cellular protection. *Hepatology* **59**, 1984–1997 (2014).
94. Yanai, H. *et al.* Conditional ablation of HMGB1 in mice reveals its protective function against endotoxemia and bacterial infection. *Proc. Natl Acad. Sci. USA* **110**, 20699–20704 (2013).
95. Vince, J. E. *et al.* Inhibitor of apoptosis proteins limit RIP3 kinase-dependent interleukin-1 activation. *Immunity* **36**, 215–227 (2012).
96. Kang, T. B., Yang, S. H., Toth, B., Kovalenko, A. & Wallach, D. Caspase-8 blocks kinase RIPK3-mediated activation of the NLRP3 inflammasome. *Immunity* **38**, 27–40 (2013).
97. Gurung, P. *et al.* FADD and caspase-8 mediate priming and activation of the canonical and noncanonical Nlrp3 inflammasomes. *J. Immunol.* **192**, 1835–1846 (2014).
98. Philip, N. H. *et al.* Caspase-8 mediates caspase-1 processing and innate immune defense in response to bacterial blockade of NF- $\kappa$ B and MAPK signaling. *Proc. Natl Acad. Sci. USA* **111**, 7385–7390 (2014).
99. Weng, D. *et al.* Caspase-8 and RIP kinases regulate bacteria-induced innate immune responses and cell death. *Proc. Natl Acad. Sci. USA* **111**, 7391–7396 (2014).
100. Wang, X. *et al.* RNA viruses promote activation of the NLRP3 inflammasome through a RIP1–RIP3–DRP1 signaling pathway. *Nature Immunol.* **15**, 1126–1133 (2014).
101. Moriwaki, K. *et al.* The necroptosis adaptor RIPK3 promotes injury-induced cytokine expression and tissue repair. *Immunity* **41**, 567–578 (2014).
102. Bossaller, L. *et al.* Cutting edge: FAS (CD95) mediates noncanonical IL-1 $\beta$  and IL-18 maturation via caspase-8 in an RIP3-independent manner. *J. Immunol.* **189**, 5508–5512 (2012).
103. Maelfait, J. *et al.* Stimulation of Toll-like receptor 3 and 4 induces interleukin-1 $\beta$  maturation by caspase-8. *J. Exp. Med.* **205**, 1967–1973 (2008).
104. Antonopoulos, C., El Sanadi, C., Kaiser, W. J., Mocarski, E. S. & Dubyak, G. R. Proapoptotic chemotherapeutic drugs induce noncanonical processing and release of IL-1 $\beta$  via caspase-8 in dendritic cells. *J. Immunol.* **191**, 4789–4803 (2013).
105. Liu, T. *et al.* Single-cell imaging of caspase-1 dynamics reveals an all-or-none inflammasome signaling response. *Cell Rep.* **8**, 974–982 (2014).
106. Cullen, S. P. *et al.* Fas/CD95-induced chemokines can serve as 'find-me' signals for apoptotic cells. *Mol. Cell* **49**, 1034–1048 (2013).
107. Krysko, D. V. *et al.* Immunogenic cell death and DAMPs in cancer therapy. *Nature Rev. Cancer* **12**, 860–875 (2012).
108. Omori, E. *et al.* TAK1 is a master regulator of epidermal homeostasis involving skin inflammation and apoptosis. *J. Biol. Chem.* **281**, 19610–19617 (2006).
109. Nenci, A. *et al.* Skin lesion development in a mouse model of incontinentia pigmenti is triggered by NEMO deficiency in epidermal keratinocytes and requires TNF signaling. *Hum. Mol. Genet.* **15**, 531–542 (2006).
110. Nenci, A. *et al.* Epithelial NEMO links innate immunity to chronic intestinal inflammation. *Nature* **446**, 557–561 (2007).
111. Kajino-Sakamoto, R. *et al.* Enterocyte-derived TAK1 signaling prevents epithelium apoptosis and the development of ileitis and colitis. *J. Immunol.* **181**, 1143–1152 (2008).
112. Panayotova-Dimitrova, D. *et al.* cFLIP regulates skin homeostasis and protects against TNF-induced keratinocyte apoptosis. *Cell Rep.* **5**, 397–408 (2013).
113. Kumari, S. *et al.* Sharpin prevents skin inflammation by inhibiting TNFR1-induced keratinocyte apoptosis. *eLife* **3**, e03422 (2014).
114. Rickard, J. *et al.* TNFR1-dependent cell death drives inflammation in Sharpin-deficient mice. *eLife* **3**, e03464 (2014).
115. Poon, I. K., Lucas, C. D., Rossi, A. G. & Ravichandran, K. S. Apoptotic cell clearance: basic biology and therapeutic potential. *Nature Rev. Immunol.* **14**, 166–180 (2014).
116. Brouckaert, G. *et al.* Phagocytosis of necrotic cells by macrophages is phosphatidylserine dependent and does not induce inflammatory cytokine production. *Mol. Biol. Cell* **15**, 1089–1100 (2004).
117. Nunes, T., Bernardazzi, C. & de Souza, H. S. Cell death and inflammatory bowel diseases: apoptosis, necrosis, and autophagy in the intestinal epithelium. *BioMed Res. Int.* **2014**, 218493 (2014).
118. Pierdomenico, M. *et al.* Necroptosis is active in children with inflammatory bowel disease and contributes to heighten intestinal inflammation. *Am. J. Gastroenterol.* **109**, 279–287 (2014).
119. Saito, N. *et al.* An annexin A1–FPR1 interaction contributes to necroptosis of keratinocytes in severe cutaneous adverse drug reactions. *Sci. Transl. Med.* **6**, 245ra95 (2014).
120. Gautheron, J. *et al.* A positive feedback loop between RIP3 and JNK controls non-alcoholic steatohepatitis. *EMBO Mol. Med.* **6**, 1062–1074 (2014).
121. Roychowdhury, S., McMullen, M. R., Pisano, S. G., Liu, X. & Nagy, L. E. Absence of receptor interacting protein kinase 3 prevents ethanol-induced liver injury. *Hepatology* **57**, 1773–1783 (2013).
122. Wickman, G. R. *et al.* Blebs produced by actin-myosin contraction during apoptosis release damage-associated molecular pattern proteins before secondary necrosis occurs. *Cell Death Differ.* **20**, 1293–1305 (2013).
123. Kazama, H. *et al.* Induction of immunological tolerance by apoptotic cells requires caspase-dependent oxidation of high-mobility group box-1 protein. *Immunity* **29**, 21–32 (2008).
124. Lüthi, A. U. *et al.* Suppression of interleukin-33 bioactivity through proteolysis by apoptotic caspases. *Immunity* **31**, 84–98 (2009).
125. Venereau, E. *et al.* Mutually exclusive redox forms of HMGB1 promote cell recruitment or proinflammatory cytokine release. *J. Exp. Med.* **209**, 1519–1528 (2012).

**Acknowledgements** We apologise to all the authors whose work we could not cite in this Review due to space limitations. M.P. acknowledges funding from the European Research Council (2012-ADG\_20120314), the German Research Foundation (SFB670, SFB829, SPP1656), the European Commission (Grants 223404 (Masterswitch) and 223151 (InflaCare)), the Deutsche Krebshilfe, the Else Kröner-Fresenius-Stiftung and the Helmholtz Alliance (PCCC). Research in the Vandenabeele unit is supported by Belgian grants (Interuniversity Attraction Poles, IAP 7/32), Flemish grants (Research Foundation Flanders, FWO G.0875.11, FWO G.0973.11 N, FWO G.0A45.12 N, FWO G.0172.12, FWO G.0787.13N, G0C3114N, FWO KAN 31528711 and Foundation against Cancer 2012-188), Gent University grants (MRP, GROUP-ID consortium) and grants from Flanders Institute for Biotechnology (VIB). P.V. holds a Methusalem grant (BOF09/01M00709) from the Flemish Government.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at [go.nature.com/Gawino](http://go.nature.com/Gawino). Correspondence should be addressed to M.P. ([pasparakis@uni-koeln.de](mailto:pasparakis@uni-koeln.de)) or P. V. ([Peter.Vandenabeele@irc.vib-ugent.be](mailto:Peter.Vandenabeele@irc.vib-ugent.be)).

# Function and information content of DNA methylation

Dirk Schübeler<sup>1,2</sup>

**Cytosine methylation is a DNA modification generally associated with transcriptional silencing. Factors that regulate methylation have been linked to human disease, yet how they contribute to malignances remains largely unknown. Genomic maps of DNA methylation have revealed unexpected dynamics at gene regulatory regions, including active demethylation by TET proteins at binding sites for transcription factors. These observations indicate that the underlying DNA sequence largely accounts for local patterns of methylation. As a result, this mark is highly informative when studying gene regulation in normal and diseased cells, and it can potentially function as a biomarker. Although these findings challenge the view that methylation is generally instructive for gene silencing, several open questions remain, including how methylation is targeted and recognized and in what context it affects genome readout.**

Methylation of nucleotides provides a molecular means to reversibly mark genomic DNA. Bacteria can methylate adenosine or cytosine to identify and degrade invading DNA and to track mismatch repair and the progress of genome duplication before cell division<sup>1,2</sup>. In eukaryotes, DNA methylation only occurs at cytosine residues. Since the discovery that *in vitro* methylated DNA is transcriptionally inactive when transfected into *Xenopus* oocytes<sup>3</sup> or cultured mammalian cells<sup>3,4</sup>, methylation has been functionally linked to gene repression. Importantly, methylation is not always essential to eukaryotic gene regulation as it is absent in many organisms. These include metazoans such as dipteran insects or *Caenorhabditis elegans*, illustrating that transcriptional changes during development do not necessarily require the organism to methylate DNA. Although DNA methylation is obligatory in many clades, its prevalence and genomic distribution varies widely, suggesting that there are distinct modes of targeting and function<sup>5–7</sup>.

Methylation of DNA can change the functional state of regulatory regions, but it does not change the Watson–Crick base pairing of cytosine. It thus presents the classic ‘epigenetic’ mark and is functionally involved in many forms of stable epigenetic repression, such as imprinting, X chromosome inactivation and silencing of repetitive DNA<sup>8</sup>. In vertebrates, heritable methylation only occurs at the CpG dinucleotide. Sequence symmetry of CpGs enables propagation of the methyl mark through cell division<sup>9</sup>; in 1975, this was proposed as a pathway for cellular memory of transcriptional states<sup>10,11</sup>. This potential for inheritance coupled with the fact that DNA methylation patterns change during development and disease<sup>8,12</sup> partially explains the interest in DNA methylation as a memory module<sup>13</sup>.

Technological advances have resulted in genomic maps of DNA methylation at unprecedented resolution, revealing that regulatory sequences are indeed unmethylated when active<sup>14–18</sup>. Despite these striking but correlative observations, our ability to correctly assign a function to the local presence of DNA methylation at particular genes has remained surprisingly limited. This Review summarizes recent advances in our understanding of the regulation and function of DNA methylation in mammals, and discusses the utility of DNA methylation as a cellular marker in basic biology and biomedicine.

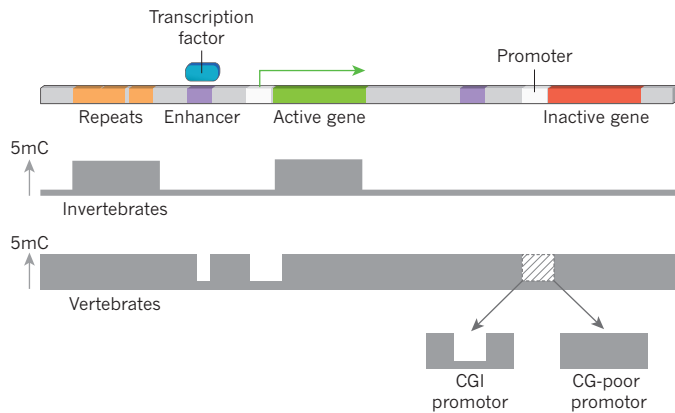
## Conservation of methylation patterns in non-vertebrates

Among the many clades that methylate their genome, vertebrates are unique in that cytosine methylation occurs throughout the entire genome

(Fig. 1). By contrast, plants and invertebrates that have been analysed so far show ‘mosaic’ methylation patterns because only specific genomic elements are targeted<sup>5</sup> (Fig. 1). More specifically, repetitive DNA and actively transcribed sequences are methylated<sup>6,7</sup>. In the case of repetitive DNA, it is evident that DNA methylation is used for repressing expression and preventing further expansion of these elements<sup>19–22</sup>. In plants, this targeting occurs through short RNAs derived from repeat transcripts that guide *de novo* methylation to this class of elements<sup>23</sup>. A similar PIWI-interacting RNA (piRNA)-guided process seems to occur in the male mammalian germ line when DNA methylation is re-established<sup>12,24</sup>. Furthermore, DNA methylation of pro-viruses has been shown to depend, in part, on the presence of repressive histone methylation at lysine 9 of histone H3 (H3K9me)<sup>25</sup>, illustrating distinct RNA and chromatin pathways that guide DNA methylation to repeats.

The second canonical target is methylation of actively transcribed genes, a process that does not seem to primarily regulate these genes<sup>6</sup>. This is prominent in organisms with mosaic DNA methylation patterns and can even occur in the absence of repeat methylation, as in the case of the invertebrate chordate *Ciona intestinalis*<sup>26,27</sup>. Despite its strong evolutionary conservation, genic methylation remains surprisingly poorly understood at both the molecular and functional level. Current models suggest that it helps to counteract the disruption of chromatin, such as nucleosome displacement, which is caused by elongating RNA polymerase<sup>8,26,27</sup>. This speculation builds on established chromatin pathways such as the marking of transcribed sequences by methylation at lysine 36 of histone H3 (H3K36me)<sup>28</sup>. Work in baker's yeast (*Saccharomyces cerevisiae*), which lacks DNA methylation, revealed that H3K36me recruits enzymes such as histone deacetylases, resulting in more densely packed chromatin. In the absence of this pathway, transcription creates a more open chromatin structure, leading to spurious activation of cryptic start sites<sup>29</sup>. Such a process might be even more relevant in the genomes of multicellular organisms, which harbour substantially larger genes. Indeed, DNA methylation has been suggested to suppress intragenic promoters in mammalian cells<sup>30</sup>, and remethylation of genes in cancer cells after treatment with an inhibitor of methylation maintenance occurs much faster at actively transcribed genes and seems to be required for their proper expression<sup>31</sup>. Notably, retroviruses preferentially integrate into actively transcribed genes<sup>32</sup>, and a process that would methylate histones and DNA and potentially silence them after genic integration thus seems

<sup>1</sup>Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, CH-4058 Basel, Switzerland. <sup>2</sup>University of Basel, Faculty of Science, Petersplatz 1, CH-4003 Basel, Switzerland.



**Figure 1 | Genomic distribution of methylated cytosine in a typical invertebrate and vertebrate genome.** The representative genomic region includes an example of an active and an inactive gene with proximal (promoter) and distal (enhancer) regulatory regions. The height of the bar indicates the relative proportion of DNA methylation (5-methylcytosine, 5mC) that is observed in each region. CpG islands (CGIs), which often overlap with promoter regions, generally remain unmethylated, whereas CG-poor promoters are methylated when not active.

plausible. Furthermore, genic methylation has been linked to splicing in mammalian cells, as subtle methylation differences were observed between introns and exons<sup>33,34</sup>. However, experimental evidence to support this link remains scarce<sup>35</sup>, and an explanation for how the elongating polymerase or splicing machinery could be affected by the comparatively small differences in methylation levels is lacking.

### Methylation in vertebrates

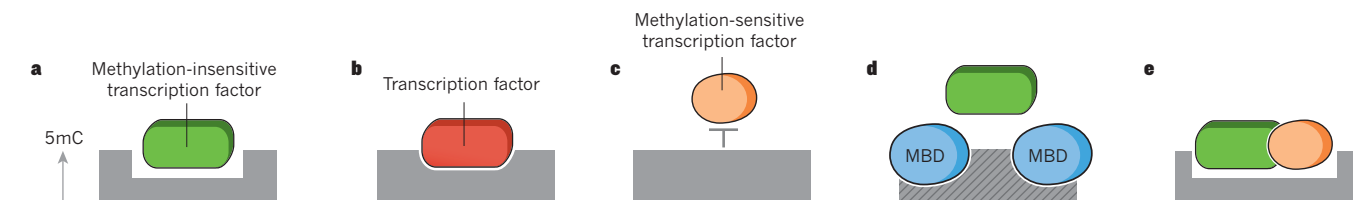
The widespread methylation of vertebrate genomes suggests that their methylation might be a default state. This causes CpG depletion over evolutionary time through inefficient base-excision repair. Although spontaneous deamination of cytosine results in uracil, a base efficiently removed by uracil-DNA glycosylase (UDG), deamination of a methylated cytosine results in thymine, a proper genomic base. The resulting mismatch frequently manifests in a C to T transition, despite the presence of thymine-DNA glycosylase (TDG) and methyl-CpG-binding domain protein 4 (MBD4): two glycosylases that are thought to target this particular mismatch within methylated CGs and lead to the removal of thymine<sup>36</sup>. Notably, a C to T transition at CpG dinucleotides is the most frequent mutation observed in human diseases and between closely related mammals, arguing that genome-wide and genic DNA methylation comes at the price of an increased mutational load. Among vertebrates, CG depletion is particularly prominent in mammalian genomes, but less obvious in those of fish or *Xenopus*<sup>37</sup>. Exceptions to this general loss of CGs reside in genomic elements termed CpG islands (CGIs)<sup>38</sup>. CGIs predominantly overlap with promoter regions and retain their expected CG content, as they generally remain unmethylated in the germ line, with notable exceptions<sup>39</sup>. Importantly, methylation of CGIs causes robust transcriptional repression and CGI methylation is

required for all reported examples of long-term mono-allelic silencing, including X inactivation<sup>40</sup> and genomic imprinting<sup>41</sup>. Although CGI methylation can result in stable repression of the linked gene (for example, see ref. 42), surprisingly few CGIs change DNA-methylation state during normal development. Prominent examples include a set of germline-specific genes that require promoter methylation for their repression in somatic cells<sup>43</sup>. Importantly, however, inactive CGI promoters do not, in general, acquire DNA methylation but become methylated at lysine 27 of histone H3 (H3K27me3), a mark set by the Polycomb system<sup>44,45</sup>. Interestingly, CGIs marked by H3K27me3 are nevertheless more susceptible to DNA methylation during differentiation and in disease states such as cancer<sup>46,47</sup>. Furthermore, CGIs harbour H3K4 methylation independently of their activity, but only in the absence of DNA methylation<sup>48</sup>. This modification can repulse *de novo* methyltransferases *in vitro* and thus is thought to be causally involved in maintaining the hypomethylated state of CGIs<sup>49</sup>.

Recent genome-wide mapping of DNA methylation at single-base resolution revealed that CG-poor regulatory regions generally acquire a low methylation state when occupied by transcription factors<sup>14–18</sup>. Such variable DNA methylation in mammals closely reflects changes in gene regulation and, as a result, methylome data provide a rich source of information about ongoing gene activity (see ‘Utility of DNA methylation’). Although the functional relevance of reduced methylation at CG-poor regulatory regions is still unclear, it is tempting to imply an instructive role for DNA methylation in distal gene regulation. This popular model assumes a generally repressive effect of DNA methylation regardless of the position and densities of CGs within a particular regulatory region. Experimental evidence does not, however, generally support this presumption. For example, several factors have been shown to bind methylated CG-poor sequences and, in turn, lead to their demethylation<sup>14,50,51</sup>. In this scenario, changes in DNA methylation occur downstream of transcription-factor binding to their target sequence, arguing against a generally instructive role. It is likely that other transcription factors are more sensitive to DNA methylation, in particular those that contain a CG in their binding motif<sup>52</sup> (Fig. 2). Potential effects might also only apply to certain binding sites such as those with lower affinity, at which DNA methylation might further reduce the likelihood of binding. Nevertheless, there is limited evidence at present that DNA methylation at CG-poor regulatory regions is generally instructive. This does not mean that methylation is irrelevant because tissue-specific deletions of, for example, DNA methyltransferases show very distinct phenotypes, including clear changes in gene expression, and mutations in these enzymes can contribute to disease<sup>12</sup>. However, it remains to be determined whether these phenotypes result from local differential methylation of regulatory regions or global perturbations, including reactivation of repeats and/or methylated CGIs.

### Setting and removing DNA methylation

In the textbook scenario, *de novo* DNA methyltransferases DNMT3A and DNMT3B<sup>53</sup> in combination with DNMT3L<sup>54</sup> establish a pattern of methylation that is then faithfully maintained through cell division by the maintenance methyltransferase DNMT1 (ref. 55) and associated proteins<sup>56</sup>. This model of stable DNA methylation propagation has recently been revised in several ways. It was long known that DNA



**Figure 2 | Potential scenarios for the interplay between cytosine methylation (shown by level of 5-methylcytosine) and transcription-factor binding.** a, A methylation-insensitive transcription factor causes reduced methylation after binding. b, A transcription factor binds specifically to the methylated state of its binding site. c, A methylation-sensitive transcription factor is blocked by

5-methylcytosine (5mC). d, Methyl-CpG-binding domain (MBD) proteins bind to the methylated state, leading to indirect repression, which probably requires high local density of CGs (shading). e, A methylation-insensitive transcription factor functions as a pioneer factor and creates a site of reduced methylation that allows a methylation-sensitive factor to bind.



methylation can be lost passively through imperfect maintenance<sup>57</sup>, but the recent discovery of the ten-eleven translocation (TET) family of proteins provided a convincing path for catalysed active demethylation in vertebrates<sup>58</sup>. The three members of this protein family have since been implicated in development, meiosis, maintenance of imprinting and stem-cell reprogramming<sup>59–64</sup>. TET proteins convert 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC), a modified base that was first described more than 30 years ago<sup>65</sup>. Further iterative oxidations catalysed by TET result in 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)<sup>66,67</sup> (Fig. 3), which can be efficiently removed by TDG<sup>68,69</sup> — a pathway referred to as ‘active’ demethylation. These modifications create an opportunity to identify sites of TET action because they can be specifically detected<sup>69–73</sup>, but their detection and quantification remains challenging owing to their low genomic abundance compared with 5mC. Current data argue that active regulatory regions and transcribed sequences show increased levels of hydroxymethylation, which is indicative of a higher turnover<sup>69,73–76</sup>. Loss of all three TET proteins leads to increased DNA methylation at enhancer regions in stem cells and subtle expression changes of linked genes<sup>77</sup>, suggesting that active demethylation could contribute to the activity state of distal regulatory regions<sup>78</sup>. Notably, active DNA demethylation could occur in most cell types because TET proteins are promiscuously expressed at varying levels in non-dividing somatic cells. Proper quantification of DNA-methylation-turnover kinetics is crucial not only for interpreting genomic maps of this mark but also for concepts of epigenetic memory that invoke DNA methylation. Memory implies stability of the modified base, which could be impaired by pathways of active demethylation. However, a conclusive picture regarding how turnover is regulated is lacking. Current models of binding and local activity for both demethylases and methyltransferases, which are largely based on *in vitro* interaction data, suggest that histone modifications and RNA–DNA hybrids have a role in repelling and recruiting these enzymes<sup>12,74,79,80</sup>. Recruitment probably constitutes just one level of regulation. For example, TET activity can be enhanced by the addition of vitamin C to the culture medium of stem cells<sup>81</sup>, suggesting that cofactors modulate enzyme activity.

### CG-specific readers

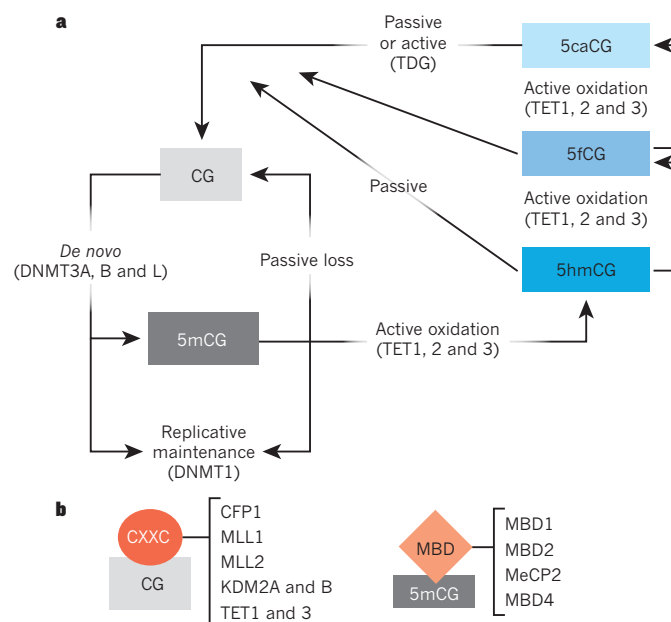
The CG dinucleotide provides distinct signals based on its methylation state that can be differentially recognized by specific protein domains. The methylated form of CG is recognized by functional MBD-containing proteins *in vitro* and *in vivo*<sup>82–84</sup>, which are thought to indirectly contribute to methylation-mediated repression<sup>85</sup> (Figs 2d, 3b). This concept of indirect repression might explain why a high local concentration of methylated CpGs, such as at methylated CGIs, do confer efficient repression. Importantly, individual deletions of genes encoding MBD proteins do not result in reactivation of methylated CGIs and thus MBD proteins are thought to function in a redundant fashion. This model awaits testing through comprehensive deletions of the MBD protein family. CXXC domains, however, can specifically recognize unmethylated CGs and can target proteins, such as CXXC finger protein 1 (CFP1)<sup>86</sup> or the histone demethylases KDM2A and KDM2B to unmethylated CGIs<sup>87,88</sup> (Fig. 3b). It has been suggested that this recruitment might help to maintain an unmethylated state similar to that seen with classic transcription factors, but evidence for this model is still missing<sup>87</sup>. Importantly, the presence of a functional CXXC domain does not necessarily predict targeting of proteins to unmethylated CGs. In the case of DNMT1, the CXXC domain seems to be used for proofreading<sup>9</sup>.

A crucial question is whether specific binders exist for cytosine oxidation derivatives, which would suggest that these bases function as distinct signals. This is a topic of active research, and several candidate proteins have been proposed<sup>74,79,89</sup>. However, the low frequency of these modifications has so far stymied efforts to convincingly show selective binding *in vivo*<sup>87</sup>. It also remains to be determined whether methylation of cytosines outside the CpG context, which seems to be a frequent event in embryonic stem cells, subsets of neurons<sup>17</sup> and oocytes<sup>90</sup>, is specifically recognized.

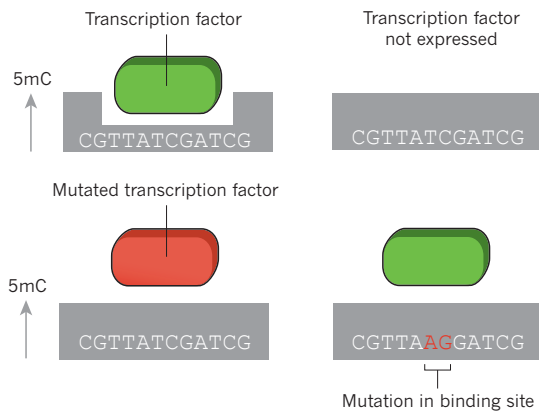
### Genetic determination and environmental influence

How much do our methylomes vary, and how can this variation be explained? Recent genome-wide association studies compared DNA sequence variation within human cohorts with changes in chromatin modifications<sup>91–93</sup> or DNA methylation<sup>94</sup>. The data showed that genetic variation explains a large part of the observed changes in histone modifications or DNA methylation<sup>91–94</sup>. The likely explanation is that mutations within regulatory regions affect binding of transcription factors, which in turn influence DNA methylation and histone modifications (Fig. 4). Alternatively, a mutation in a transcription factor could impair its ability to bind a specific sequence, leading to increased methylation. This transcription-factor dependency is similar to that observed during development, in which expression of cell-type-specific transcription factors coincides with reduced methylation of their binding sites (Fig. 4). These genome-wide correlations are also supported by gene-insertion experiments showing that local DNA sequence can be the primary determinant of DNA methylation state<sup>95,96</sup>. Indeed a role for transcription factors in establishing DNA methylation states was observed 20 years ago<sup>50,97,98</sup>. Thus several lines of evidence reinforce the notion that the underlying DNA sequence, through its recognition by transcription factors, seems to account for a substantial part of the observed DNA methylation and chromatin patterns and their variation between individuals or cell types.

How transcription factors influence chromatin and DNA methylation and which factors are instructive or sensitive to chromatin states remains largely unexplored (Fig. 2). These are crucial questions, because understanding this genetic and epigenetic crosstalk will be essential to correctly interpret and assign function to patterns of DNA methylation in development and disease. It should help in identifying sites that are more susceptible to sporadic change and environmental exposure. Indeed, differences in DNA methylation are observed in isogenic subjects, such as identical



**Figure 3 | Setting, erasing and recognizing cytosine methylation.** **a**, Different methylation states of the CG dinucleotide and the enzymatic pathways that set, maintain and erase the mark. The pathways leading from the oxidized forms to the unmethylated state are under debate<sup>74</sup>. DNMT, DNA methyltransferase; TDG, thymine–DNA glycosylase. **b**, A subset of CXXC-domain-containing proteins are listed that can specifically bind to the unmethylated CG dinucleotide and could potentially reinforce the unmethylated state or recruit regulatory proteins to unmethylated CG islands. Methyl–CpG-binding domain (MBD) proteins specifically bind to the methylated CG with little or no further sequence sensitivity, potentially mediating transcriptional repression, which would be strongest in methylated CG islands. Readers of oxidized forms are not shown owing to ongoing debate about proposed candidates<sup>87</sup>.



**Figure 4 | Potential DNA sequence determinants of cytosine methylation at CG-poor regions.** In a simplified model, transcription-factor binding causes reduced methylation at its binding site. Loss of expression of the respective transcription factor in development or disease will cause increased methylation. Mutations to the transcription factor that affect its binding preference will influence genomic methylation patterns. Mutation in the DNA binding site will abolish binding even in a cell expressing the transcription factor, indicating how genetic variation can result in methylation differences between individuals.

twins, but the actual level is under debate<sup>99</sup>. Furthermore, developing mouse embryos display methylation changes after severe undernutrition *in utero*<sup>100</sup>. The question remains as to whether these differences occur downstream of perturbed gene activity or are causally involved in the resulting phenotype. In the case of the undernutrition model, the metabolic phenotype can even propagate across generations, despite the lack of differences in DNA methylation<sup>100</sup>. Generally, although a role for DNA methylation in trans-generational inheritance is clearly appealing, there is limited evidence for this pathway in mammals (reviewed in ref. 101). At the same time, stably inherited epialleles clearly exist in plants<sup>101</sup>. One explanation for this difference is that mammals show waves of DNA-methylation gain and loss in the germ line and early development, leading to the erasure of most acquired methylation. This important aspect of DNA methylation biology is not the focus of this Review and has recently been discussed in detail<sup>12,13,24</sup>.

### New disease links

Perturbations of DNA methylation patterns are frequently observed in disease, particularly in cancer. These perturbations include methylation of CGI promoters for tumour suppressor genes, implying a functional role<sup>8</sup>. At the time of the initial observations, support for this model in the form of recurring mutations, in tumours, of proteins that regulate DNA methylation was missing. This changed with the discovery of the enzymatic activity of TET proteins — TET2 mutations have been linked to many myeloid malignancies<sup>102,103</sup> — and through the observation that DNMT3A is one of the most frequently mutated genes in acute myeloid leukaemia (AML) (mutated in 25% of adults with the disease)<sup>104</sup>. Although the actual mechanism in both cases remains to be determined, it is striking that TET2 and DNMT3A loss of function seem to be primary events. Mutations in DNMT3A are pre-leukaemic and can occur in blood stem cells, but only in combination with subsequent ‘driver mutations’ do they lead to leukaemia<sup>105,106</sup>.

In fact, recurrent mutations of many genes encoding chromatin components have now been identified in diverse cancer types<sup>107</sup>. It is unclear how these proteins actually contribute to disease, but one hypothesis is that chromatin perturbations make the regulatory landscape more vulnerable to subsequent mutations<sup>105</sup>. Such concepts of genetic buffering by chromatin pathways have already been proposed based on genetic interaction screens between hypomorphic mutants, for example in *C. elegans*<sup>108</sup>. In this study, mutations in chromatin modifiers enhanced the phenotype of mutations in many different transcription factors. One interpretation of these findings is that

chromatin pathways globally contribute to the stability of transcriptional regulation rather than directing activity to specific genes. It is tempting to speculate that DNA methylation might also contribute in this way to genome regulation.

### Utility of DNA methylation

Although the function of DNA methylation at CG-poor regulatory regions remains unclear, it is evident that patterns of this mark have high information content about the ongoing activity of transcription factors and, thus, can help to identify cell-type-specific aspects of gene regulation<sup>109</sup>. For this reason, DNA methylation profiles can provide insight into many aspects of biology, but also function as biomarkers in medicine. It is important to keep in mind that the value of a biomarker is only defined by its ability to predict disease state or treatment response regardless of whether the measured parameter is causally involved in the disease. Part of the appeal of DNA methylation is the feasibility of the analysis. Bisulphite conversion of unmethylated cytosines followed by DNA sequencing requires no live cells and limited amounts of DNA, which can even be low quality, making DNA methylation less sensitive to specimen handling compared with RNA or proteins<sup>110</sup>. As a result, bisulphite sequencing, which, importantly, cannot distinguish between 5mC and 5hmC, can even be conducted on DNA isolated from small amounts of fixed tissue<sup>111</sup>. Any primary sample from which DNA can be sequenced should be suitable for bisulphite sequencing, a technique readily performed in any laboratory that is set-up for genome sequencing. Given that technical issues do not represent a significant hurdle, the crucial question is how much medically relevant information can be obtained from such DNA methylation analysis. Potential pitfalls include cellular heterogeneity within disease samples, confounding genetic variation between individuals or intrinsic variability, which could limit the utility of methylation measurements.

Nevertheless, in the case of medulloblastoma, a detailed analysis of methylation levels in regulatory regions enabled identification of disease- and tumour-subtype-specific changes<sup>112</sup>. These changes reflect not only transcription-factor activity but also putative new markers for disease states<sup>112</sup>. The aforementioned study adds to a list of examples for which DNA methylation can assist in identifying tumour subtypes<sup>8</sup>.

Moreover, in most disease settings, cells from the affected tissue cannot be easily obtained, and it is questionable whether those that are available will be informative (for example, blood cells). Several studies have now shown that non-affected tissue can be informative<sup>113</sup>, even leading to treatment decisions<sup>114</sup> and stratification of clinical cohorts<sup>115</sup>. Further insights into the functionality of DNA methylation will come from ongoing large cohort studies, leading towards a more comprehensive view of whether and how disease phenotypes can be associated with specific DNA methylation patterns<sup>116</sup>. The link between gene activity and DNA methylation, as well as the accessibility of this base modification, justifies further exploration. One intriguing possibility is that DNA methylation might be prognostic of disease onset or personal risk in combination with genetic predisposition and environmental exposure.

The potential use of DNA methylation is not limited to biomedical applications. Measurements at the resolution of individual molecules and cells promise to substantially advance our understanding of regulation in biology. They provide more quantitative information and reveal heterogeneity within the cell population or dynamic changes at the level of individual cells<sup>117</sup>. Bisulphite conversion followed by high-throughput sequencing represents a single-molecule measure and has already been successfully applied in stem-cell biology to define population heterogeneity and the degree of pluripotency after induced reprogramming<sup>118,119</sup>. For example, subpopulations of mouse embryonic stem cells that differ in the level of the transcription factor Rex1 can be separated by their DNA methylation patterns<sup>120</sup>. Based on chemical inhibition, DNA methylation was argued to directly contribute to the switch between these populations<sup>120</sup>. Studies that tackle questions of cell identity and origin will further benefit from the recent development of approaches that allow genome-wide measurements at the level of single cells<sup>121</sup>.

Tracking methylation patterns over time can also reveal insights into

the fidelity of DNA methylation maintenance. A comparison of the heterogeneity of DNA methylation at specific genomic loci suggested that human stem cells in culture undergo high turnover of DNA methylation. By contrast, fibroblast cells seem to show more stable inheritance of frequent sporadic changes, implying that somatic cells show more faithful maintenance and thus pass on sporadic changes in DNA methylation<sup>122</sup>. If the same applies *in vivo*, high-coverage DNA methylation measurements from single cells should provide information on lineage trees. This approach, when applied to an *in vitro* model of cellular transformation, has revealed insights into the kinetics of progressive changes in regions that are commonly hypermethylated in cancer<sup>123</sup>. Although it remains to be seen whether this principle is relevant to carcinogenesis *in vivo*, it highlights how DNA methylation can be used to track cellular states over time.

## Outlook

The emerging picture is that genomic DNA methylation in mammals reflects, to a large extent, cell-intrinsic regulation encoded within the DNA sequence in the form of CG density and binding motifs for transcription factors. This genetic dependency is very reminiscent of histone modifications, which are also an integral part of the activation of regulatory regions and the process of transcription. It explains why chromatin and DNA modifications are informative indicators of underlying regulatory activity, but it does not reveal what their actual impact is on the regulation of individual genes. This is likely to be highly contextual at the level of DNA sequence and binding factors, as shown by the efficient repression of CGIs by DNA methylation, which might represent the exception rather than the rule because dynamics in DNA methylation occur elsewhere in the genome. This exemplifies our limited understanding of how local levels of DNA methylations are set, turned over by demethylation and, in turn, recognized by transacting factors. A better understanding of this regulation is required to define these potential functions and determine how misregulation contributes to disease. ■

Received 17 September; accepted 19 November 2014.

- Arber, W. & Dussoix, D. Host specificity of DNA produced by *Escherichia coli*. I. Host controlled modification of bacteriophage lambda. *J. Mol. Biol.* **5**, 18–36 (1962).
- Wion, D. & Casades, J. N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nature Rev. Microbiol.* **4**, 183–192 (2006).
- Vardimon, L., Kressmann, A., Cedar, H., Maechler, M. & Doerfler, W. Expression of a cloned adenovirus gene is inhibited by *in vitro* methylation. *Proc. Natl Acad. Sci. USA* **79**, 1073–1077 (1982).
- Stein, R., Razin, A. & Cedar, H. *In vitro* methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proc. Natl Acad. Sci. USA* **79**, 3418–3422 (1982).
- Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nature Rev. Genet.* **9**, 465–476 (2008).
- Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010).
- Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA* **107**, 8689–8694 (2010).
- Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Rev. Genet.* **13**, 484–492 (2012).
- Song, J., Rechkoblit, O., Bestor, T. H. & Patel, D. J. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science* **331**, 1036–1040 (2011).
- Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14**, 9–25 (1975).
- Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187**, 226–232 (1975).
- Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Rev. Genet.* **14**, 204–220 (2013).
- Lee, H. J., Hore, T. A. & Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14**, 710–719 (2014).
- Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
- Hon, G. C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Genet.* **45**, 1198–1206 (2013).
- Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
- Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).
- Hodges, E. *et al.* Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell* **44**, 17–28 (2011).
- Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).
- Selker, E. U. Epigenetic phenomena in filamentous fungi: useful paradigms or repeat-induced confusion? *Trends Genet.* **13**, 296–301 (1997).
- Lippman, Z. *et al.* Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**, 471–476 (2004).
- Walsh, C. P., Chaillat, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genet.* **20**, 116–117 (1998).
- This study showed that absence of DNA methylation in the developing mouse embryo leads to reactivation of retroviruses.**
- Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Rev. Genet.* **11**, 204–220 (2010).
- Hill, P. W., Amouroux, R. & Hajkova, P. DNA demethylation, Tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: an emerging complex story. *Genomics* **104**, 324–333 (2014).
- Liu, S. *et al.* Setdb1 is required for germline development and silencing of H3K9me3-marked endogenous retroviruses in primordial germ cells. *Genes Dev.* **28**, 2041–2055 (2014).
- Suzuki, M. M., Kerr, A. R., De Sousa, D. & Bird, A. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res.* **17**, 625–631 (2007).
- Simmen, M. W. *et al.* Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* **283**, 1164–1167 (1999).
- Keogh, M. C. *et al.* Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* **123**, 593–605 (2005).
- Carrozza, M. J. *et al.* Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123**, 581–592 (2005).
- Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010).
- Yang, X. *et al.* Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**, 577–590 (2014).
- Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2**, e234 (2004).
- Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res.* **23**, 1256–1269 (2013).
- Chodavarapu, R. K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466**, 388–392 (2010).
- Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011).
- Cortázar, D. *et al.* Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature* **470**, 419–423 (2011).
- Long, H. K. *et al.* Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife* **2**, e00348 (2013).
- Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209–213 (1986).
- Smallwood, S. A. *et al.* Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature Genet.* **43**, 811–814 (2011).
- Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genet.* **33** (Suppl.), 245–254 (2003).
- Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).
- This work established that imprinted gene expression requires DNA methylation.**
- Velasco, G. *et al.* Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. *Proc. Natl Acad. Sci. USA* **107**, 9281–9286 (2010).
- Borgel, J. *et al.* Targets and dynamics of promoter DNA methylation during early mouse development. *Nature Genet.* **42**, 1093–1100 (2010).
- Tanay, A., O'Donnell, A. H., Damelin, M. & Bestor, T. H. Hyperconserved CpG domains underlie Polycomb-binding sites. *Proc. Natl Acad. Sci. USA* **104**, 5521–5526 (2007).
- Lynch, M. D. *et al.* An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. *EMBO J.* **31**, 317–329 (2012).
- Ohm, J. E. *et al.* A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature Genet.* **39**, 237–242 (2007).
- Mohn, F. *et al.* Lineage-specific polycomb targets and *de novo* DNA methylation define restriction and potential of neuronal progenitors. *Mol. Cell* **30**, 755–766 (2008).
- Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genet.* **39**, 457–466 (2007).
- Ooi, S. K. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to *de novo* methylation of DNA. *Nature* **448**, 714–717 (2007).
- Han, L., Lin, I. G. & Hsieh, C. L. Protein binding protects sites on stable episomes and in the chromosome from *de novo* methylation. *Mol. Cell Biol.* **21**, 3416–3424 (2001).
- Kress, C., Thomassin, H. & Grange, T. Active cytosine demethylation triggered by a nuclear receptor involves DNA strand breaks. *Proc. Natl Acad. Sci. USA* **103**, 11112–11117 (2006).
- Liu, Y., Zhang, X., Blumenthal, R. M. & Cheng, X. A common mode of recognition for methylated CpG. *Trends Biochem. Sci.* **38**, 177–183 (2013).
- Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. *Cell* **99**, 247–257 (1999).
- Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. & Bestor, T. H. Dnmt3L and the establishment of maternal genomic imprints. *Science* **294**, 2536–2539 (2001).



55. Yoder, J. A. & Bestor, T. H. A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. *Hum. Mol. Genet.* **7**, 279–284 (1998).
56. Sharif, J. *et al.* The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450**, 908–912 (2007).
57. Chen, T., Ueda, Y., Dodge, J. E., Wang, Z. & Li, E. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Mol. Cell. Biol.* **23**, 5594–5605 (2003).
58. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
- This work established TET enzymes as active DNA demethylases and 5hmC as their product.**
59. Gu, T. P. *et al.* The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477**, 606–610 (2011).
60. Quivoron, C. *et al.* TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer Cell* **20**, 25–38 (2011).
61. Ficiz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398–402 (2011).
62. Yamaguchi, S. *et al.* Tet1 controls meiosis by regulating meiotic gene expression. *Nature* **492**, 443–447 (2012).
63. Dawlaty, M. M. *et al.* Combined deficiency of Tet1 and Tet2 causes epigenetic abnormalities but is compatible with postnatal development. *Dev. Cell* **24**, 310–323 (2013).
64. Wu, H. *et al.* Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473**, 389–393 (2011).
65. Penn, N. W., Suwalski, R., O'Riley, C., Bojanowski, K. & Yura, R. The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *Biochem. J.* **126**, 781–790 (1972).
66. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
67. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
68. Zhang, L. *et al.* Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nature Chem. Biol.* **8**, 328–330 (2012).
69. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
70. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343–348 (2011).
71. Pastor, W. A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394–397 (2011).
72. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
73. Song, C. X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678–691 (2013).
74. Pastor, W. A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature Rev. Mol. Cell Biol.* **14**, 341–356 (2013).
75. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S. E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.* **12**, R54 (2011).
76. Feldmann, A. *et al.* Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet.* **9**, e1003994 (2013).
77. Lu, F., Liu, Y., Jiang, L., Yamaguchi, S. & Zhang, Y. Role of Tet proteins in enhancer activity and telomere elongation. *Genes Dev.* **28**, 2103–2119 (2014).
78. Hon, G. C. *et al.* 5mC Oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. *Mol. Cell* **56**, 286–297 (2014).
79. Wu, H. & Zhang, Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell* **156**, 45–68 (2014).
80. Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I. & Chedin, F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell* **45**, 814–825 (2012).
81. Blaschke, K. *et al.* Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature* **500**, 222–226 (2013).
82. Baubec, T., Ivanek, R., Lienert, F. & Schubeler, D. Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell* **153**, 480–492 (2013).
83. Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L. & Bird, A. P. Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* **58**, 499–507 (1989).
- This study reported the identification of the first protein that specifically binds the methylated CG dinucleotide.**
84. Hendrich, B. & Bird, A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.* **18**, 6538–6547 (1998).
85. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31**, 89–97 (2006).
86. Thomson, J. P. *et al.* CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**, 1082–1086 (2010).
87. Hahn, M. A., Szabo, P. E. & Pfeifer, G. P. 5-Hydroxymethylcytosine: a stable or transient DNA modification? *Genomics* **104**, 304–323 (2014).
88. Long, H. K., Blackledge, N. P. & Klose, R. J. ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochem. Soc. Trans.* **41**, 727–740 (2013).
89. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146–1159 (2013).
90. Tomizawa, S. *et al.* Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development* **138**, 811–820 (2011).
91. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
92. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–747 (2013).
93. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
94. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2**, e00523 (2013).
95. Lienert, F. *et al.* Identification of genetic elements that autonomously determine DNA methylation states. *Nature Genet.* **43**, 1091–1097 (2011).
96. Krebs, A., Dessus-Babus, S., Burger, L. & Schubeler, D. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *eLife* **3**, e04094 (2014).
97. Brandeis, M. *et al.* Sp1 elements protect a CpG island from *de novo* methylation. *Nature* **371**, 435–438 (1994).
98. Macleod, D., Charlton, J., Mullins, J. & Bird, A. P. Sp1 sites in the mouse *aprt* gene promoter are required to prevent methylation of the CpG island. *Genes Dev.* **8**, 2282–2292 (1994).
99. Bell, J. T. & Spector, T. D. DNA methylation studies using twins: what are they telling us? *Genome Biol.* **13**, 172 (2012).
100. Radford, E. J. *et al.* *In utero* undernourishment perturbs the adult sperm methylome and intergenerational metabolism. *Science* **345**, 1255903 (2014).
101. Heard, E. & Martienssen, R. A. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell* **157**, 95–109 (2014).
102. Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839–843 (2010).
103. Ono, R. *et al.* LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23). *Cancer Res.* **62**, 4075–4080 (2002).
104. Ley, T. J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363**, 2424–2433 (2010).
105. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).
106. Corces-Zimmerman, M. R., Hong, W. J., Weissman, I. L., Medeiros, B. C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl Acad. Sci. USA* **111**, 2548–2553 (2014).
107. Plass, C. *et al.* Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nature Rev. Genet.* **14**, 765–780 (2013).
108. Lehner, B., Crombie, C., Tischler, J., Fortunato, A. & Fraser, A. G. Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nature Genet.* **38**, 896–903 (2006).
109. Burger, L., Gaidatzis, D., Schubeler, D. & Stadler, M. B. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res.* **41**, e155 (2013).
110. Heyn, H. & Esteller, M. DNA methylation profiling in the clinic: applications and challenges. *Nature Rev. Genet.* **13**, 679–692 (2012).
111. Schillebeekx, M. *et al.* Laser capture microdissection-reduced representation bisulfite sequencing (LCM-RRBS) maps changes in DNA methylation associated with gonadectomy-induced adrenocortical neoplasia in the mouse. *Nucleic Acids Res.* **41**, e116 (2013).
112. Hovestadt, V. *et al.* Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* **510**, 537–541 (2014).
113. Rakyian, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nature Rev. Genet.* **12**, 529–541 (2011).
114. Weller, M. *et al.* MGMT promoter methylation in malignant gliomas: ready for personalized medicine? *Nature Rev. Neurol.* **6**, 39–51 (2010).
115. Jacquemont, S. *et al.* Epigenetic modification of the FMR1 gene in fragile X syndrome is associated with differential response to the mGluR5 antagonist AFQ056. *Sci. Transl. Med.* **3**, 64ra61 (2011).
116. Bock, C. Analysing and interpreting DNA methylation data. *Nature Rev. Genet.* **13**, 705–719 (2012).
117. Junker, J. P. & van Oudenaarden, A. Every cell is special: genome-wide studies add a new dimension to single-cell biology. *Cell* **157**, 8–11 (2014).
118. Ma, H. *et al.* Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature* **511**, 177–183 (2014).
119. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
120. Singer, Z. S. *et al.* Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Mol. Cell* **55**, 319–331 (2014).
121. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods* **11**, 817–820 (2014).
122. Shipony, Z. *et al.* Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**, 115–119 (2014).
123. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nature Genet.* **44**, 1207–1214 (2012).

**Acknowledgements** I apologize to colleagues, whose work I could not cite or only discuss in a limited context owing to space limitations. I thank in particular P. Ginno and further T. Baubec, M. Lorincz and N. Thomae for critical input on the manuscript. Work in my laboratory is supported by the Novartis research foundation, the European Union (NoE EpiGeneSys FP7-HEALTH-2010-257082 and the Blueprint consortium FP7-282510), the European Research Council (EpiGePlas), the SNF Sinergia programme and the Swiss Initiative in Systems Biology (RTD Cell Plasticity).

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The author declares no competing financial interests. Readers are welcome to comment on the online version of this paper at [go.nature.com/hdj5ba](http://go.nature.com/hdj5ba). Correspondence should be addressed to D.S. ([dirk@fmi.ch](mailto:dirk@fmi.ch)).

# The African Genome Variation Project shapes medical genetics in Africa

Deepti Gurdasani<sup>1,2\*</sup>, Tommy Carstensen<sup>1,2\*</sup>, Fasil Tekola-Ayele<sup>3\*</sup>, Luca Pagani<sup>1,4\*</sup>, Ioanna Tachmazidou<sup>1\*</sup>, Konstantinos Hatzikotoulas<sup>1</sup>, Savita Karthikeyan<sup>1,2</sup>, Louise Iles<sup>1,2,5</sup>, Martin O. Pollard<sup>1</sup>, Ananyo Choudhury<sup>6</sup>, Graham R. S. Ritchie<sup>1,7</sup>, Yali Xue<sup>1</sup>, Jennifer Asimit<sup>1</sup>, Rebecca N. Nsubuga<sup>8</sup>, Elizabeth H. Young<sup>1,2</sup>, Cristina Pomilla<sup>1,2</sup>, Katja Kivinen<sup>1</sup>, Kirk Rockett<sup>9</sup>, Anatoli Kamali<sup>8</sup>, Ayo P. Doumatey<sup>3</sup>, Gershon Asiki<sup>8</sup>, Janet Seeley<sup>8</sup>, Fatoumatta Sisay-Joof<sup>10</sup>, Muminatou Jallow<sup>10</sup>, Stephen Tollman<sup>11,12</sup>, Ephrem Mekonnen<sup>13</sup>, Rosemary Ekong<sup>14</sup>, Tamiru Oljira<sup>15</sup>, Neil Bradman<sup>16</sup>, Kalifa Bojang<sup>10</sup>, Michele Ramsay<sup>6,17,18</sup>, Adebawale Adeyemo<sup>3</sup>, Endashaw Bekele<sup>19</sup>, Ayesha Motala<sup>20</sup>, Shane A. Norris<sup>21</sup>, Fraser Pirie<sup>20</sup>, Pontiano Kaleebu<sup>8</sup>, Dominik Kwiatkowski<sup>1,2</sup>, Chris Tyler-Smith<sup>1</sup>§, Charles Rotimi<sup>3</sup>§, Eleftheria Zeggini<sup>1</sup>§ & Manjinder S. Sandhu<sup>1,2</sup>§

**Given the importance of Africa to studies of human origins and disease susceptibility, detailed characterization of African genetic diversity is needed. The African Genome Variation Project provides a resource with which to design, implement and interpret genomic studies in sub-Saharan Africa and worldwide. The African Genome Variation Project represents dense genotypes from 1,481 individuals and whole-genome sequences from 320 individuals across sub-Saharan Africa. Using this resource, we find novel evidence of complex, regionally distinct hunter-gatherer and Eurasian admixture across sub-Saharan Africa. We identify new loci under selection, including loci related to malaria susceptibility and hypertension. We show that modern imputation panels (sets of reference genotypes from which unobserved or missing genotypes in study sets can be inferred) can identify association signals at highly differentiated loci across populations in sub-Saharan Africa. Using whole-genome sequencing, we demonstrate further improvements in imputation accuracy, strengthening the case for large-scale sequencing efforts of diverse African haplotypes. Finally, we present an efficient genotype array design capturing common genetic variation in Africa.**

Globally, human populations show structured genetic diversity as a result of geographical dispersion, selection and drift. Understanding this variation can provide insights into evolutionary processes that shape both human adaptation and variation in disease susceptibility<sup>1</sup>. Although the Hapmap Project<sup>2</sup> and the 1000 Genomes Project<sup>3</sup> have greatly enhanced our understanding of genetic variation globally, the characterization of African populations remains limited. Other efforts examining African genetic diversity have been limited by variant density and sample sizes in individual populations<sup>4</sup>, or have focused on isolated groups, such as hunter gatherers (HG)<sup>5,6</sup>, limiting relevance to more widespread populations across Africa.

The African Genome Variation Project (AGVP) is an international collaboration that expands on these efforts by systematically assessing genetic diversity among 1,481 individuals from 18 ethno-linguistic groups from sub-Saharan Africa (SSA) (Fig. 1 and Supplementary Methods Tables 1 and 2) with the HumanOmni2.5M genotyping array and whole-genome sequences (WGS) from 320 individuals (Supplementary

Methods Table 2). Importantly, the AGVP has evolved to help develop local resources for public health and genomic research, including strengthening research capacity, training, and collaboration across the region. We envisage that data from this project will provide a global resource for researchers, as well as facilitate genetic studies in Africa<sup>7</sup>.

## Population structure in SSA

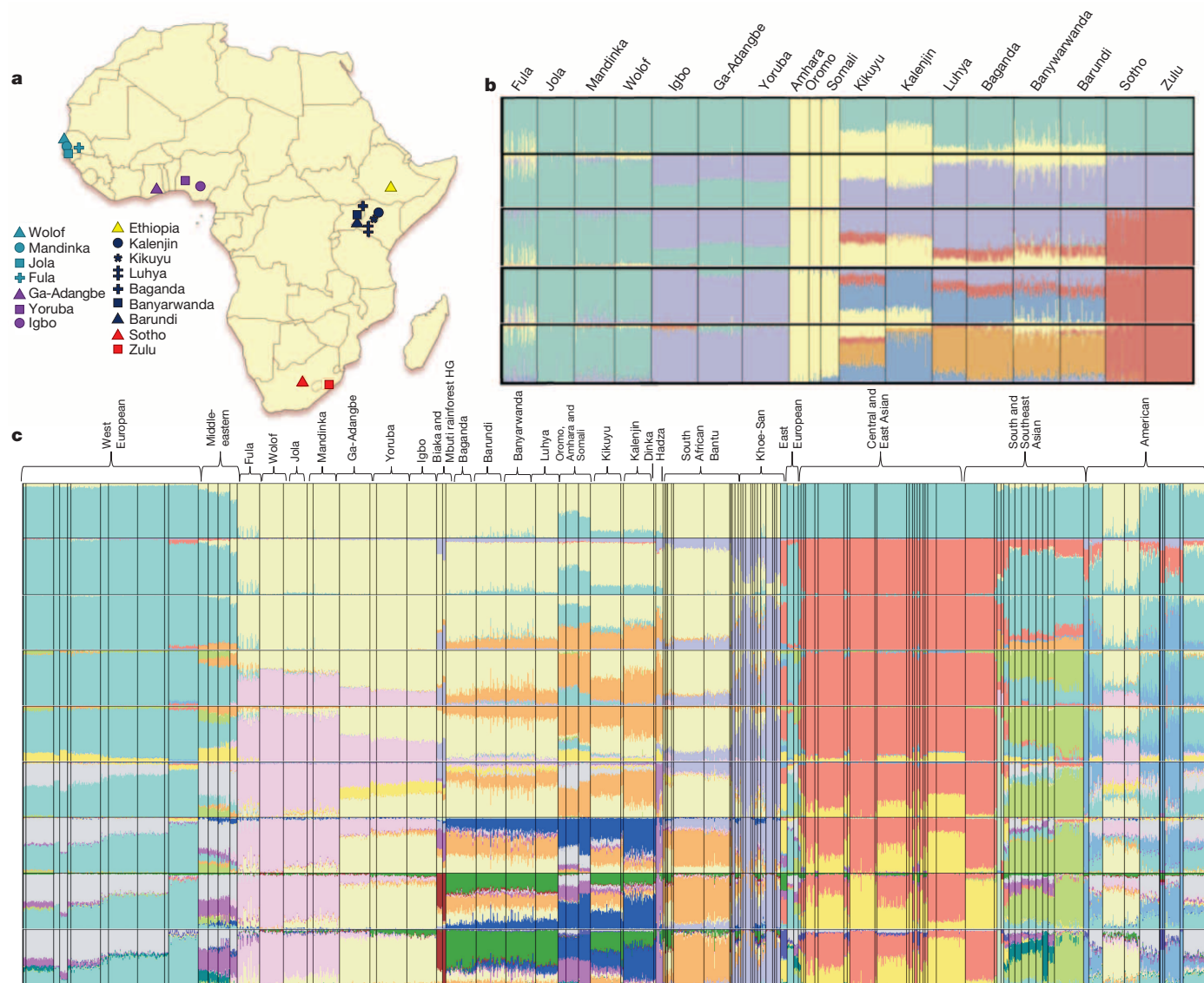
On examining ~2.2 million variants, we found modest differentiation among SSA populations (mean pairwise  $F_{ST}$  0.019) (Supplementary Methods and Supplementary Table 1). Differentiation among the Niger-Congo language groups—the predominant linguistic grouping across Africa was noted to be modest (mean pairwise  $F_{ST}$  0.009) (Supplementary Table 1), providing evidence for the ‘Bantu expansion’—a recent population expansion and movement throughout SSA originating in West Africa around 3,000 to 5,000 years ago<sup>8</sup>.

We identified 29.8 million single-nucleotide polymorphisms (SNPs) from Ethiopian, Zulu and Bagandan WGS (Extended Data Fig. 1 and

<sup>1</sup>Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK. <sup>2</sup>Department of Public Health and Primary Care, University of Cambridge, 2 Wort's Causeway, Cambridge, CB1 8RN, UK. <sup>3</sup>Centre for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, 12 South Drive, MSC 5635, Bethesda, Maryland 20891-5635, USA. <sup>4</sup>Department of Biological, Geological and Environmental Sciences, University of Bologna, Via Selmi 3, 40126 Bologna, Italy. <sup>5</sup>Department of Archaeology, University of York, King's Manor, York YO1 7EP, UK. <sup>6</sup>Sydney Brenner Institute of Molecular Bioscience (SBIMB), University of the Witwatersrand, The Mount, 9 Jubilee Road, Parktown 2193, Johannesburg, Gauteng, South Africa. <sup>7</sup>Vertebrate Genomics, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. <sup>8</sup>Medical Research Council/Uganda Virus Research Institute, Plot 51-57 Nakiwogo Road, Uganda. <sup>9</sup>Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Headington, Oxford OX3 7BN, UK. <sup>10</sup>Medical Research Council Unit, Atlantic Boulevard, Serrekunda, PO Box 273, Banjul, The Gambia. <sup>11</sup>Medical Research Council/Wits Rural Public Health and Health Transitions Unit, School of Public Health, Education Campus, 27 St Andrew's Road, Parktown 2192, Johannesburg, Gauteng, South Africa. <sup>12</sup>INDEPTH Network, 38/40 Mensah Wood Street, East Legon, PO Box KD 213, Kanda, Accra, Ghana. <sup>13</sup>Institute of Biotechnology, Addis Ababa University, Entoto Avenue, Arat Kilo, 16087 Addis Ababa, Ethiopia. <sup>14</sup>Department of Genetics Evolution and Environment, University College, London, Gower Street, London WC1E 6BT, UK. <sup>15</sup>University of Haramaya, Department of Biology, PO Box 138, Dire Dawa, Ethiopia. <sup>16</sup>Henry Stewart Group, 28/30 Little Russell Street, London WC1A 2HN, UK. <sup>17</sup>Division of Human Genetics, National Health Laboratory Service, C/O Hospital and de Korte Streets, Braamfontein 2000, Johannesburg, South Africa. <sup>18</sup>School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Braamfontein 2000, Johannesburg, South Africa. <sup>19</sup>Department of Microbial, Cellular and Molecular Biology, College of Natural Sciences, Arat Kilo Campus, Addis Ababa University, PO Box 1176, Addis Ababa, Ethiopia. <sup>20</sup>Department of Diabetes and Endocrinology, University of KwaZulu-Natal, 719 Umbilo Road, Congella, Durban 4013, South Africa. <sup>21</sup>Department of Paediatrics, University of Witwatersrand, 7 York Road, Parktown 2198, Johannesburg, Gauteng, South Africa.

\*These authors contributed equally to this work.

§These authors jointly supervised this work.



**Figure 1 | Populations studied in the AGVP.** **a**, 18 African populations studied in the AGVP including 2 populations from the 1000 Genomes Project. (The term ‘Ethiopia’ encompasses the Oromo, Amhara and Somali ethno-linguistic groups.) **b**, **c**, ADMIXTURE analysis of these 18 populations alone ( $n = 1,481$ ) (**b**) and in a global context ( $n = 3,904$ ) (**c**). Each colour represents a different ancestral cluster, with clusters 2–6 represented along the y-axis in **b**

Supplementary Methods). A substantial proportion of unshared (11–23%) and novel (16–24%) variants were observed, with the highest proportion among Ethiopian populations (Extended Data Fig. 1). The high proportion of unshared variation among populations recapitulates the need for large-scale sequencing across Africa, including among genetically divergent populations.

We used principal component analysis to explore relationships among AGVP populations (Extended Data Figs 2–5, Supplementary Figs 1 and 2). PC1 appeared to represent a cline extending from West and East African populations towards Ethiopian populations, possibly suggesting Eurasian gene flow, while PC2 separated West African and South/East African populations (Extended Data Fig. 2). Inclusion of the 1000 Genomes Project, North African and Khoe-San (Khoisan) populations in principal component analysis (Extended Data Figs 3–5, and Supplementary Figs 1 and 2) suggested possible HG ancestry among southern Niger-Congo groups—highlighted by clustering towards the Khoe-San, in addition to confirming a cline towards Eurasian populations. ‘Unsupervised’ (that is, without including known information on individual ancestry)

and clusters 2–18 represented in **c**.  $K = 6$  and  $K = 18$  were the most likely clusters on ADMIXTURE analysis. ADMIXTURE analysis suggests substructure between North, East, West and South Africa. Studying these populations in the context of Eurasian and African HG populations suggest extensive Eurasian and HG admixture across Africa.

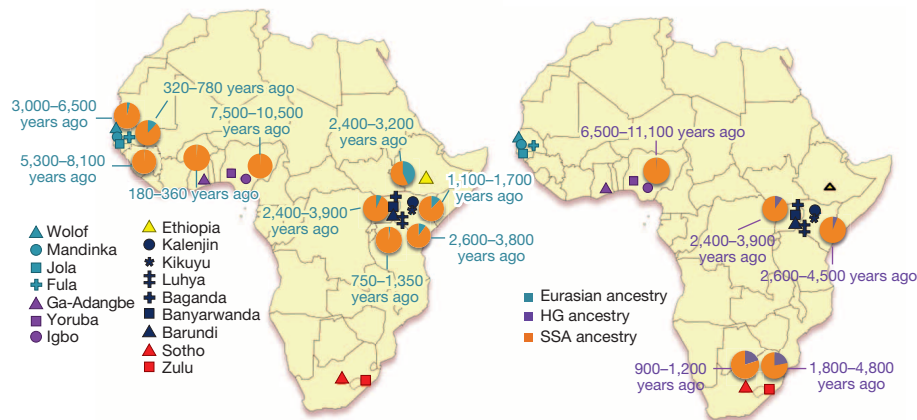
ADMIXTURE<sup>9</sup> (<https://www.genetics.ucla.edu/software/admixture/>) analysis including the 1000 Genomes Project and Human Origins data sets (Fig. 1), also supported evidence for substantial Eurasian and HG ancestry in SSA (Fig. 1 and Extended Data Fig. 6).

To assess the effect of gene flow on population differentiation in SSA, we masked Eurasian ancestry across the genome (Supplementary Methods and Supplementary Note 6). This markedly reduced population differentiation, as measured by a decline in mean pairwise  $F_{ST}$  from 0.021 to 0.015 (Supplementary Note 6), suggests that Eurasian ancestry has a substantial impact on differentiation among SSA populations. We speculate that residual differentiation between Ethiopian and other SSA populations after masking Eurasian ancestry (pairwise  $F_{ST} = 0.027$ ) may be a remnant of East African diversity pre-dating the Bantu expansion<sup>10</sup>.

### Population admixture in SSA

Formal tests for admixture (the three population test or  $f_3$  statistic)<sup>11</sup>, confirmed widespread Eurasian and HG admixture in SSA (Supplementary Tables 2 and 3). Quantification of admixture (Supplementary Table 4,





**Figure 2 | Dating and proportion of Eurasian and HG admixture among African populations.** The proportion and distribution of Eurasian and HG admixture among different populations across Africa, with approximate

dating of admixture using MALDER (code was provided by J. Pickrell; see Supplementary Information).

Supplementary Methods and Supplementary Notes 3 and 4) indicated substantial Eurasian ancestry in many African populations (ranging from 0% to 50%), with the greatest proportion in East Africa (Fig. 2 and Supplementary Table 4). Similarly, HG admixture ranged from 0% to 23%, being greatest among Zulu and Sotho (Fig. 2 and Supplementary Table 5).

We found evidence for historically complex and regionally distinct admixture with multiple HG and Eurasian populations across SSA (Fig. 2 and Supplementary Note 5). Specifically, ancient Eurasian admixture was observed in central West African populations (Yoruba; ~7,500–10,500 years ago), old admixture among Ethiopian populations (~2,400–3,200 years ago) consistent with previous reports<sup>10,12</sup>, and more recent complex admixture in some East African populations (~150–1,500 years ago) (Fig. 2, Extended Data Fig. 7 and Supplementary Note 5). Our finding of ancient Eurasian admixture corroborates findings of non-zero Neanderthal ancestry in Yoruba, which is likely to have been introduced through Eurasian admixture and back migration, possibly facilitated by greening of the Sahara desert during this period<sup>13,14</sup>.

We also find evidence for complex and regionally distinct HG admixture across SSA (Fig. 2, Extended Data Figs 7 and Supplementary Note 5), with ancient gene flow (~9,000 years ago) among Igbo and more recent admixture in East and South Africa (multiple events ranging from 100 years ago to 3,000 years ago), broadly consistent with historical movements reflecting the Bantu expansion. An exploration of the likeliest sources of admixture in our data suggested that HG admixture in Igbo was most closely represented by modern day Khoe-San populations rather than by rainforest HG populations (Supplementary Note 5). Given limited archaeological and linguistic evidence for the presence of Khoe-San populations in West Africa, this extant HG admixture might represent ancient populations, consistent with the presence of mass HG graves from the early Holocene period comprising skeletons with distinct morphological features<sup>15</sup>, and with evidence of HG rock art dating to this period in the western Sahara<sup>16,17</sup>. In East Africa, our analyses suggested that Mbuti rainforest HG populations most closely represented ancient HG mixing populations (Supplementary Note 5), with admixture dating to ~3,000 years ago, suggesting that HG ancestry here is likely to be older than previously reported<sup>18</sup>. The primary source of HG admixture in Zulu and Sotho populations was from Khoe-San populations (Fig. 2 and Supplementary Note 5), consistent with linguistic assimilation of click consonants among these populations.

### Positive selection in SSA

We examined highly differentiated SNPs between European and African populations, as well as among African populations to gain insights into loci that may have undergone selection in response to local adaptive forces (Supplementary Methods). To account for confounding due to

Eurasian admixture, we also conducted analyses after masking Eurasian ancestry (Supplementary Methods and Supplementary Note 6).

On examining locus-specific Europe–Africa differentiation, enrichment of loci known to be under positive selection was observed among the most differentiated sites ( $P = 1.4 \times 10^{-31}$ ). Furthermore, there was statistically significant enrichment for gene variants among these, indicating that this differentiation is unlikely to have arisen purely from random drift ( $P = 0.0002$ ). Additionally, we found no evidence for background selection as the primary driver of differentiation among these loci (Supplementary Note 7).

In addition to genes known to be under positive selection (for example, *SLC24A5*, *SLC45A2* and *OCA2*<sup>19,20</sup>, *LARGE*<sup>21</sup> and *CYP3A4/5*) (Supplementary Fig. 3), we found evidence of differentiation in novel gene regions, including one implicated in malaria (for chemokine receptor 1, *CR1*) (Extended Data Fig. 8). *CR1* carries the Knops blood group antigens and has previously been implicated in malaria susceptibility<sup>22</sup> and severity<sup>23</sup>, with evidence suggesting positive selection in malaria-endemic regions<sup>24</sup> (Extended Data Fig. 8). We also identified highly differentiated variants within genes involved in osmoregulation (*ATP1A1* and *AQP2*) (Extended Data Fig. 8). Deregulation of *AQP2* expression and loss-of-function mutations in *ATP1A1* have been associated with essential and secondary hypertension, respectively<sup>25,26</sup>. Climatic adaptive changes in these gene regions could potentially provide a biological basis for the high burden of hypertension and differences in salt sensitivity observed in SSA<sup>27</sup>.

In contrast, overall differentiation among African populations was modest (maximum masked  $F_{ST} = 0.19$ ) (Supplementary Fig. 4) and only 56/1,237 sites remained in the tail distribution after masking (Supplementary Methods, Supplementary Table 6). This suggests that a large proportion of differentiation observed among African populations could be due to Eurasian admixture, rather than adaptation to selective forces (Supplementary Note 6). Genes known to be under selection were notably enriched among the most differentiated loci after masking of Eurasian ancestry ( $P = 2.3 \times 10^{-16}$ ). Among the 56 loci robust to Eurasian ancestry masking (Supplementary Table 6), we identified several loci known to be under selection (Extended Data Fig. 8), including a highly differentiated variant (rs1378940) in the *CSK* gene region implicated in hypertension in genome-wide association studies (GWAS)<sup>28</sup>. The major allele of rs1378940 among Africans was in complete linkage disequilibrium with the risk allele of the GWAS SNP rs1378942 (ref. 29), with the frequency of this allele highly correlated with latitude ( $r = -0.67$ ), providing support for local adaptation in response to temperature as a possible mechanism for hypertension (Supplementary Fig. 5)<sup>30,31</sup>.

Comparing populations residing in endemic and non-endemic infectious disease regions (Supplementary Methods), we identified several

loci associated with infectious disease susceptibility and severity. As well as the known sickle-cell locus related to malaria, this approach identified additional signals for genes potentially under selection, including the *PKLR* region<sup>32</sup>, *RUNX3*<sup>33</sup>, the haptoglobin locus, *CD163*<sup>34</sup>, *IL10*<sup>35,36</sup>, *CFH*, and the *CD28-ICOS-CTLA4* locus (Supplementary Table 7 and Extended Data Fig. 8)<sup>37</sup>. Similar comparisons for Lassa fever identified the known *LARGE* gene, as well as candidates associated with viral entry and immune response, including in the Histocompatibility Leukocyte Antigen region, *DC-SIGN/DC-SIGNR*<sup>38</sup> (also known as *CD209/CLEC4M*), *RNASEL*, *CXCR6*, *IFIH1*<sup>39</sup> and *OAS2/3* regions (Supplementary Table 7). For trypanosomiasis, we identified *APOL1*<sup>40</sup>, as well as several loci implicated in immune response and binding to trypanosoma, including *FAS*, *FASLG*<sup>41,42</sup>, *IL23R*<sup>43</sup>, *SIGLEC6* and *SIGLEC12* (Supplementary Table 7)<sup>44</sup>. For trachoma, we identified signals in *ABCA1* and *CXCR6*, which may be important for the growth of the parasite and host immune response, respectively (Supplementary Table 7)<sup>45,46</sup>.

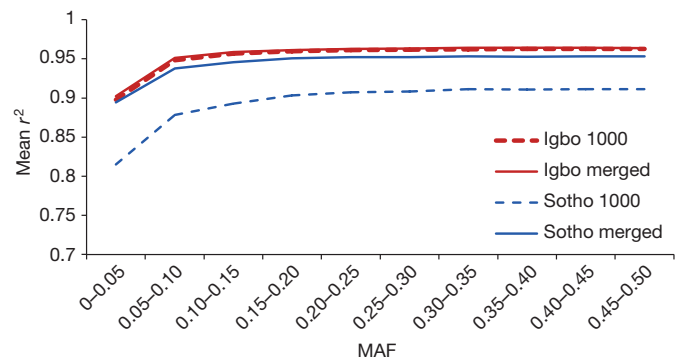
### Designing medical genetics studies in Africa

To inform the design of genomic studies in Africa, we addressed the following questions: (1) How well do current genotype arrays perform in African populations using existing reference panels for imputation? (2) Can these genotype arrays and reference panels identify and fine-map association signals in populations across Africa? (3) Can we improve imputation accuracy in African populations using a new African reference panel? and (4) What are the most cost-effective designs for large-scale GWAS in Africa?

The 1000 Genomes Project phase I integrated panel provided reasonably accurate imputation into the Illumina Omni 2.5M array in all populations (Supplementary Note 10). However, imputation accuracy was lower among Sotho, Zulu and Afro-Asiatic populations, possibly reflecting poor representation of some African haplotypes (including Khoe-San haplotypes) within the 1000 Genomes Project panel. These findings suggest that improvements in imputation accuracy across diverse population groups may require larger and more diverse reference panels.

We assessed the reproducibility and potential for fine-mapping association signals within Africa and globally at several disease susceptibility loci (Supplementary Methods, Supplementary Table 8 and Extended Data Fig. 9). Current genotype arrays and imputation panels allowed for identification of relevant association signals at most loci across populations in SSA, demonstrating that association signals are reproducible across populations in SSA (Extended Data Fig. 9 and Supplementary Figs 7–18). African populations are likely to provide better fine-mapping resolution around the causal locus (Supplementary Table 8). We highlight one example here: the sickle-cell anaemia locus (*HBB*)<sup>47</sup>, which is under positive selection owing to the protection the sickle cells confer against severe malaria. This locus showed marked heterogeneity in association signals across populations, reflecting different linkage disequilibrium patterns and allele frequencies among populations in SSA (Supplementary Figs 9 and 10). This pattern is probably the result of independent selection sweeps at this locus in different parts of Africa, leading to differences in hitchhiking rare haplotypes that attained high frequencies among different populations<sup>48</sup>. This suggests that these signatures are recent and occurred during or after the Bantu expansion, consistent with the hypothesis that the advent of agriculture and increased malaria transmission may have resulted in increased selection pressure<sup>49</sup>. However, in contrast to previous reports<sup>47</sup>, we show that association signals even at such highly differentiated loci can be captured with dense genotype data using existing reference panels for imputation, despite individual population groups not being fully represented in these. This suggests that, instead of large-scale population-specific sequencing across Africa, what is needed is a broad sequencing approach, targeted at capturing widespread haplotype diversity.

To assess the utility of a larger and more diverse African reference panel for imputation, we generated a panel integrating the 1000 Genomes Project phase I and AGVP WGS panels (Supplementary Methods and Supplementary Note 9). Using this integrated panel, we observed marked



**Figure 3 | Improvement in imputation accuracy with the AGVP WGS panel.** The substantial improvement in imputation accuracy in some populations (Sotho), compared to minimal improvement in others (Igbo) with the addition of the AGVP WGS reference panel to the 1000 Genomes Project phase I reference panel ('merged') suggests poor representation of some haplotypes (for example, Khoe-San haplotypes in Sotho) in the 1000 Genomes Project reference panel alone ('1000').  $r^2$  is the correlation coefficient, representing the correlation between imputed and genotyped data, on masking each genotyped variant during imputation. MAF, minor allele frequency.

improvements in imputation accuracy across the whole range of the allele frequency spectrum in specific populations poorly represented by the 1000 Genomes Project panel (Fig. 3 and Supplementary Note 11). These findings suggest that even common haplotypes in some SSA populations may not be sufficiently captured by existing panels, limiting our power to examine associations of common variants with disease. Importantly, given the specificity of the improvement in imputation accuracy, we infer that targeted sequencing of divergent populations representing a broad spectrum of haplotypes across Africa, including HG and North/East African haplotypes, rather than widespread population sequencing is likely to provide a more efficient strategy to improve imputation accuracy and a practicable GWAS framework in Africa.

We compared the utility of existing chip designs (2.5M Illumina) and ultralow-coverage WGS designs (0.5×, 1×, 2× coverage) to determine the optimal design for African GWAS. Sensitivity for common variation was >90% at all sequencing depths (Supplementary Note 12). Examining the effective sample size for a fixed budget<sup>50</sup>, we found the effective sample size was greater for all ultralow-coverage WGS and chip array designs compared with 4× WGS. When computational costs were accounted for (Supplementary Note 12), the HumanOmni2.5M array provided the greatest effective sample size supporting the development and large-scale use of efficient genotype arrays in Africa, where these have been underutilized.

We therefore sought to evaluate a potential chip design to tag common variation across a wider range of African populations (Supplementary Note 13). Importantly, we show that an array with one million genetic variants could capture >80% of common variation (minor allele frequency >5%) across the genome (Extended Data Fig. 10). These analyses suggest that designing a pan-African genotype array to effectively capture common genetic variation across Africa is feasible, and could greatly facilitate large-scale genomic studies in Africa.

### Discussion

The marked haplotype diversity within Africa has important implications for the design of large-scale medical genomics studies across the region, as well as studies of population history and evolution. In this context, the AGVP is a resource that will facilitate a broad range of genomic studies in Africa and globally.

Although Africa is the most genetically diverse region in the world, we provide evidence for relatively modest differentiation among populations representing the major sub-populations in SSA, consistent with recent population movement and expansion across the region beginning around 5,000 years ago—the Bantu expansion<sup>8</sup>. Although the history

of the Bantu expansion is probably complex, assessments of population admixture can provide new insights. We note historically complex and regionally distinct admixture with multiple HG and Eurasian populations across SSA, including ancient HG and Eurasian ancestry in West and East Africa and more recent complex HG admixture in South Africa. As well as explaining genetic differentiation among modern populations in SSA, these admixture patterns provide genetic evidence for early back-to-Africa migrations, the possible existence of extant HG populations in western Africa—compatible with archaeological evidence<sup>15</sup>, and patterns of gene flow consistent with the Bantu expansion, including genetic assimilation of populations resident across the region.

This admixture also has important implications for the assessment of differentiation and positive selection in Africa. Accounting for these elements, we have identified loci under positive selection that are linked with hypertension, malaria, and other pathogens. This provides a proof-of-concept for the ability of geographically widespread genetic data within Africa to identify loci under selection related to diverse environments.

Our evidence for the broad transferability of genetic association signals and their statistical refinement has important implications for medical genetic research in Africa. Importantly, we highlight that such studies are feasible and can be enabled through the development of more efficient genotype arrays and diverse WGS reference panels for accurate imputation of common variation. In this context, we describe a framework for a new pan-African genotype array that could directly facilitate large-scale genomic studies in Africa.

A critical next step is the large-scale deep sequencing of multiple and diverse populations across Africa, which should be integrated with ancient DNA data. This would enable us to identify and understand signals of ancient admixture, patterns of historical population movements, and to provide a comprehensive resource for medical genomic studies in Africa.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 15 July; accepted 23 October 2014.**

**Published online 3 December 2014.**

- Botigué, L. R. *et al.* Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc. Natl Acad. Sci. USA* **110**, 11791–11796 (2013).
- The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
- The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
- Schlebusch, C. M. *et al.* Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* **338**, 374–379 (2012).
- Jarvis, J. P. *et al.* Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genet.* **8**, e1002641 (2012).
- The H3Africa Consortium. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).
- de Filippo, C., Bostoen, K., Stoneking, M. & Pakendorf, B. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proc. R. Soc. Lond. B* **279**, 3256–3263 (2012).
- Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- Pagani, L. *et al.* Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *Am. J. Hum. Genet.* **91**, 83–96 (2012).
- Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
- Pickrell, J. K. *et al.* Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl Acad. Sci. USA* **111**, 2632–2637 (2014).
- Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
- Kuper, R. & Kropelin, S. Climate-controlled Holocene occupation in the Sahara: motor of Africa's evolution. *Science* **313**, 803–807 (2006).
- Sereno, P. C. *et al.* Lakeside cemeteries in the Sahara: 5000 years of holocene population and environmental change. *PLoS ONE* **3**, e2995 (2008).
- The Bradshaw Foundation. *The Origin of the Prehistoric Rock Art Artists* <http://www.bradshawfoundation.com/giraffe/artists.php> (2014).
- Tilman, L.-E. Rock art in African Highlands, Ennedi Highlands, Chad—Artists and Herders in a Lifeworld on the Margins. In *Atlas of Cultural and Environmental Change in Arid Africa* [http://www.academia.edu/1580718/Rock\\_art\\_in\\_African\\_Highlands\\_Ennedi\\_Highlands\\_Chad\\_-\\_Artists\\_and\\_Herders\\_in\\_a\\_Lifeworld\\_on\\_the\\_Margins](http://www.academia.edu/1580718/Rock_art_in_African_Highlands_Ennedi_Highlands_Chad_-_Artists_and_Herders_in_a_Lifeworld_on_the_Margins) (Heinrich Barth Institute, 2007).
- Patin, E. *et al.* The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nature Commun.* **5**, 3163, <http://dx.doi.org/10.1038/ncomms4163> (2014).
- Norton, H. L. *et al.* Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol. Biol. Evol.* **24**, 710–722 (2007).
- Edwards, M. *et al.* Association of the OCA2 polymorphism His615Arg with melanin content in east Asian populations: further evidence of convergent evolution of skin pigmentation. *PLoS Genet.* **6**, e1000867 (2010).
- Andersen, K. G. *et al.* Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Phil. Trans. R. Soc. Lond. B* **367**, 868–877 (2012).
- Eid, N. A. *et al.* Candidate malaria susceptibility/protective SNPs in hospital and population-based studies: the effect of sub-structuring. *Malar. J.* **9**, 119 (2010).
- Panda, A. K. *et al.* Complement receptor 1 variants confer protection from severe malaria in Odisha, India. *PLoS ONE* **7**, e49420 (2012).
- Kosoy, R. *et al.* Evidence for malaria selection of a CR1 haplotype in Sardinia. *Genes Immun.* **12**, 582–588 (2011).
- Beuschlein, F. *et al.* Somatic mutations in ATP1A1 and ATP2B3 lead to aldosterone-producing adenomas and secondary hypertension. *Nature Genet.* **45**, 440–444, <http://dx.doi.org/10.1038/ng.2550> (2013).
- Graffe, C. C., Bech, J. N., Lauridsen, T. G., Vase, H. & Pedersen, E. B. Abnormal increase in urinary aquaporin-2 excretion in response to hypertonic saline in essential hypertension. *BMC Nephrol.* **13**, 15 (2012).
- Young, J. H. *et al.* Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet.* **1**, e82 (2005).
- Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
- Tabara, Y. *et al.* Common variants in the ATP2B1 gene are associated with susceptibility to hypertension: the Japanese Millennium Genome Project. *Hypertension* **56**, 973–980 (2010).
- Hong, K. W. *et al.* Genetic variations in ATP2B1, CSK, ARSG and CSMD1 loci are related to blood pressure and/or hypertension in two Korean cohorts. *J. Hum. Hypertens.* **24**, 367–372 (2010).
- Levy, D. *et al.* Genome-wide association study of blood pressure and hypertension. *Nature Genet.* **41**, 677–687 (2009).
- Machado, P. *et al.* Malaria: looking for selection signatures in the human PKLR gene region. *Br. J. Haematol.* **149**, 775–784 (2010).
- Band, G. *et al.* Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genet.* **9**, e1003509 (2013).
- Kusi, K. A. *et al.* Levels of soluble CD163 and severity of malaria in children in Ghana. *Clin. Vaccine Immunol.* **15**, 1456–1460 (2008).
- Zhang, G. *et al.* Interleukin-10 (IL-10) polymorphisms are associated with IL-10 production and clinical malaria in young children. *Infect. Immun.* **80**, 2316–2322 (2012).
- Wilson, J. N. *et al.* Analysis of IL10 haplotypic associations with severe malaria. *Genes Immun.* **6**, 462–466 (2005).
- Jacobs, T., Graefe, S. E., Niknafs, S., Gaworski, I. & Fleischer, B. Murine malaria is exacerbated by CTLA-4 blockade. *J. Immunol.* **169**, 2323–2329 (2002).
- Shimojima, M., Stroher, U., Ebihara, H., Feldmann, H. & Kawaoka, Y. Identification of cell surface molecules involved in dystroglycan-independent Lassa virus cell entry. *J. Virol.* **86**, 2067–2078 (2012).
- Fumagalli, M. *et al.* Population genetics of IFI1: ancient population structure, local selection, and implications for susceptibility to type 1 diabetes. *Mol. Biol. Evol.* **27**, 2555–2566 (2010).
- Ko, W. Y. *et al.* Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. *Am. J. Hum. Genet.* **93**, 54–66 (2013).
- Lopes, M. F. *et al.* Increased susceptibility of Fas ligand-deficient *gld* mice to *Trypanosoma cruzi* infection due to a Th2-biased host immune response. *Eur. J. Immunol.* **29**, 81–89 (1999).
- Martins, G. A. *et al.* Fas-FasL interaction modulates nitric oxide production in *Trypanosoma cruzi*-infected mice. *Immunology* **103**, 122–129 (2001).
- Ribeiro, C. M. *et al.* Trypanosomiasis-induced Th17-like immune responses in carp. *PLoS ONE* **5**, e13012 (2010).
- Crocker, P. R., Paulson, J. C. & Varki, A. Siglecs and their roles in the immune system. *Nature Rev. Immunol.* **7**, 255–266 (2007).
- Cox, J. V., Naher, N., Abdelrahman, Y. M. & Belland, R. J. Host HDL biogenesis machinery is recruited to the inclusion of *Chlamydia trachomatis*-infected cells and regulates chlamydial growth. *Cell. Microbiol.* **14**, 1497–1512 (2012).
- Natividad, A. *et al.* Human conjunctival transcriptome analysis reveals the prominence of innate defense in *Chlamydia trachomatis* infection. *Infect. Immun.* **78**, 4895–4911 (2010).
- Jallow, M. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genet.* **41**, 657–665 (2009).
- Teo, Y. Y. *et al.* Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res.* **19**, 1849–1860 (2009).
- Hedrick, P. W. Population genetics of malaria resistance in humans. *Heredity* **107**, 283–304 (2011).
- Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genet.* **44**, 631–635 (2012).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This project was funded in part by the Wellcome Trust (grant number WT077383/Z/05/Z), The Wellcome Trust Sanger Institute (grant number WT098051), the Bill and Melinda Gates Foundation, the Foundation for the National Institutes of Health (grant number 566), and the UK Medical Research Council (grant



numbers G0901213-92157, G0801566, G0600718 and MR/K013491/1). We also acknowledge the National Institute for Health Research Cambridge Biomedical Research Centre and the Wellcome Trust Cambridge Centre for Global Health Research. We are very grateful to J. Pickrell for sharing human origins data and MALDER code, and for useful input on interpretations of these analyses. We also thank E. Garrison for his suggestions on using Genome-in-a-bottle sets ([ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant\\_calls/NIST/README.NIST.v2.18.txt](ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/NIST/README.NIST.v2.18.txt)) for validation of whole-genome sequencing data. We also thank the African Partnership for Chronic Disease Research (APCDR) for providing a network to support this study as well as a repository for deposition of curated data. Sample collections from South Africa were funded by The South African Sugar Association, Servier South Africa and The Victor Daitz Foundation. The Kenyan samples were collected by D. Ngare of Moi University, Eldoret, Kenya, as part of the Africa America Diabetes Mellitus (AADM) study and the International HapMap project (D. Ngare, who is now deceased, was a great supporter of genomics in Africa, as exemplified by his leadership in engaging the Luhya and Maasai communities for the HapMap project). The Igbo samples were collected by J. Oli of the University of Nigeria, Enugu, Nigeria. The Ga-Adangbe samples were collected by the laboratories of A. Amoah of the University of Ghana, Accra, Ghana, and J. Acheampong of the University of Science and Technology, Kumasi, Ghana. Support for the AADM study is provided by the National Institute on Minority Health and Health Disparities, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and the National Human Genome Research Institute (NHGRI). The Gambian samples were collected by M. Jallow and colleagues at the MRC Unit, The Gambia and form part of the MalariaGEN Consortial Resource. This research was supported in part by the Intramural Research Program of the Center for Research on Genomics and Global Health (CRGGH; grant number Z01HG200362) and by the MRC Centre for Genomics and Global Health. D.G. was funded by the Cambridge Commonwealth Scholarship. We thank the 1000 Genomes Project for sharing genotype data that were analysed as part of this project. We also thank all study participants who contributed to this study.

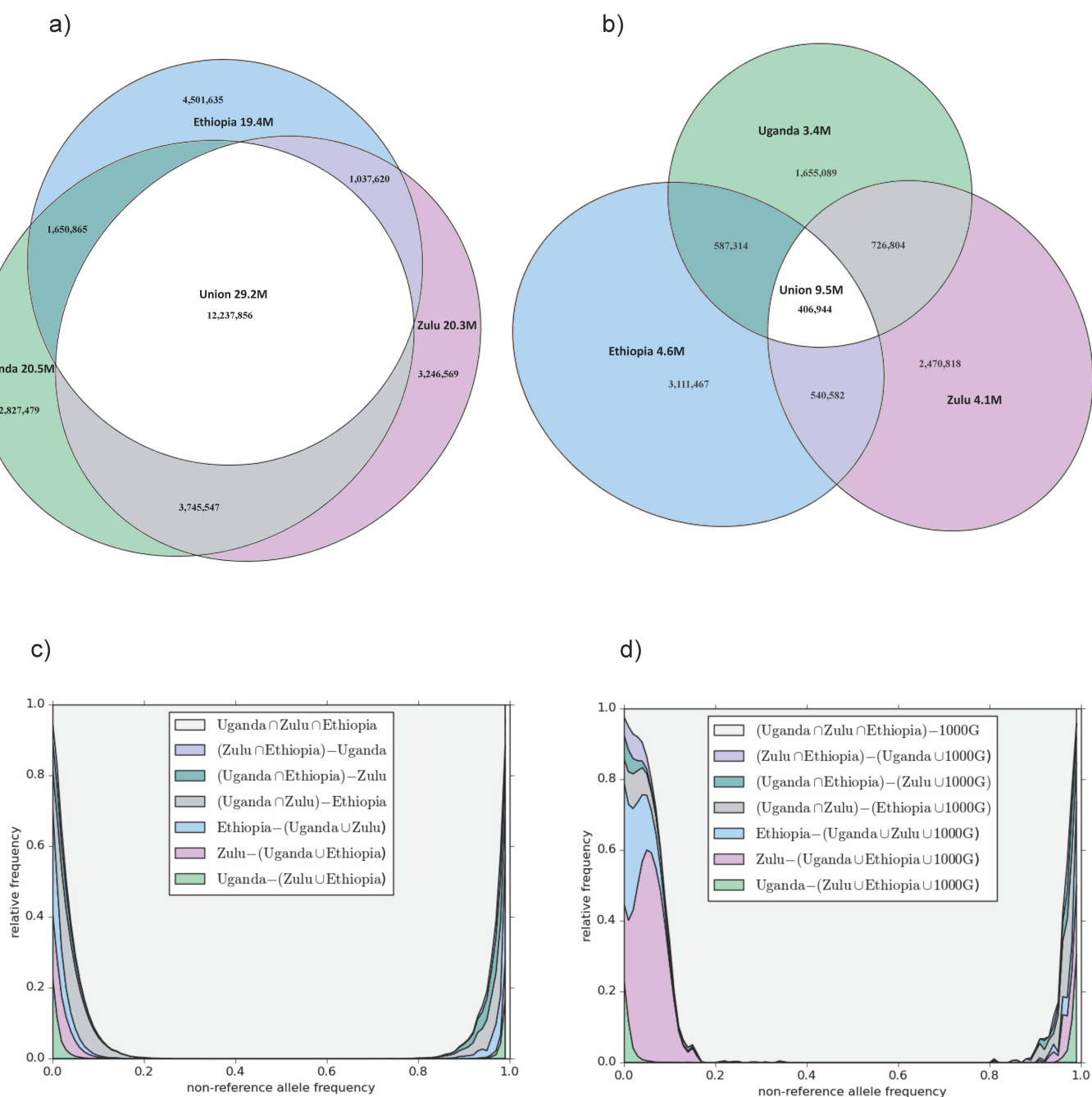
**Author Contributions** Overall project coordination: D.G., C.P., M.S.S. (Project Chair), E.H.Y. and E.Z. coordinated the project. Analysis and writing: C.P. coordinated sample collation, genotyping, quality control and data generation for the study. J.A., T.C., D.G. and C.P. carried out quality control and curation of data. R.N. and Y.X. undertook quality control for MalariaGEN and Ethiopian population sets respectively. M.O.P. carried out quality control and bam (sequencing reads file format) improvement of sequence data

at all depths. T.C. curated and generated all sequence data, and carried out comparisons with genotype array data and with higher coverage data. D.G. carried out the population structure and admixture analyses. A.C., D.G., S.K. and L.P. carried out analysis of positive selection and population differentiation. L.P. and I.T. carried out analysis of linkage disequilibrium decay. T.C., K.H. and I.T. carried out imputation-based analyses. T.C. developed an efficient tagging algorithm and carried out analysis for coverage of tagging variants for the design of the African genotype array. D.G. and F.T.-A. carried out fine mapping analyses. C.R., M.S.S., C.T.-S. and E.Z. critically appraised and commented on the manuscript. D.G., T.C., L.P. and M.S.S. prepared the manuscript and the Supplementary Information. C.P. and L.I. contributed to the writing of the Supplementary Information. All authors commented on the interpretation of results, and reviewed and approved the final manuscript. Management, fieldwork, laboratory analyses and coordination of contributing cohorts: K.B., M.J., K.K., D.K., K.R. and F.S.-J. (the Gambian cohorts—MalariaGEN); G.A., P.K., A.K., M.S.S. and J.S. (The General Population Cohort Study); A.M. and F.P. (the South African Zulu cohort); A.A., A.P.D., C.R. and F.T.-A. (the Kenyan, Ghanaian and Nigerian cohorts); A.C., S.N., M.R. and S.T. (the South African Sotho cohort); and E.B., N.B., R.E., E.M., T.O., L.P. and C.T. (the Ethiopian cohort).

**Author Information** The ADMIXTURE code is available at <https://www.genetics.ucla.edu/software/admixture/download.html>. The MALDER software is available from J. Pickrell ([jkpickrell@nygenome.org](mailto:jkpickrell@nygenome.org)). All other source code can be obtained by contacting D.G. ([dg11@sanger.ac.uk](mailto:dg11@sanger.ac.uk)). See Supplementary Methods for details. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.S.S. ([ms23@sanger.ac.uk](mailto:ms23@sanger.ac.uk)), E.Z. ([eleftheria@sanger.ac.uk](mailto:eleftheria@sanger.ac.uk)), C.R. ([rotimic@mail.nih.gov](mailto:rotimic@mail.nih.gov)) and C.T.-S. ([cts@sanger.ac.uk](mailto:cts@sanger.ac.uk)).

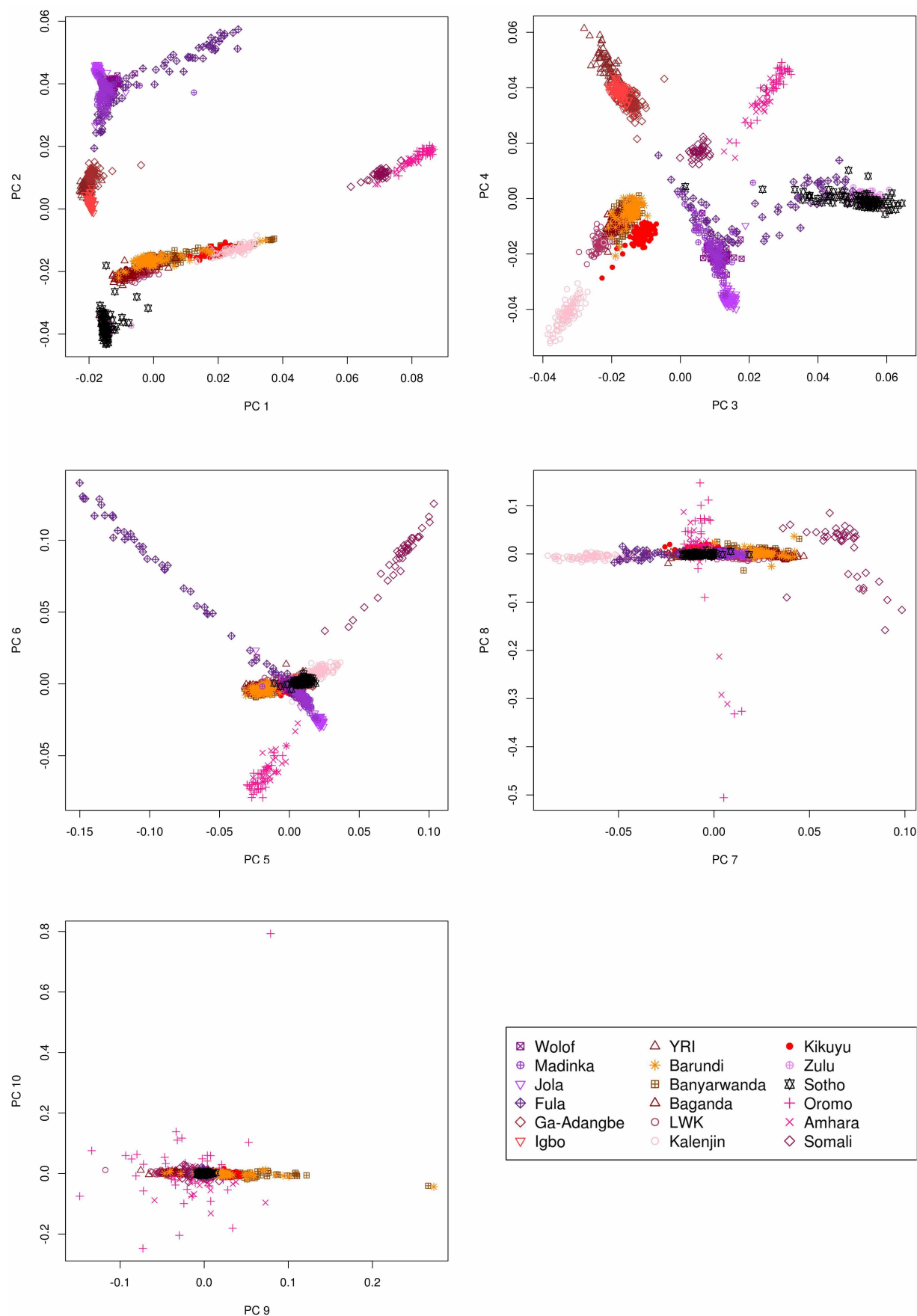


This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>



**Extended Data Figure 1 | Allele sharing between sequenced populations in the AGVP. a,** The overlap of SNPs between 4×WGS data from Zulu, Ugandan and Ethiopian individuals (subsampling to 100 samples each). **b,** The overlap of novel variants (those not in the 1000 Genomes Project phase I integrated call set, '1000G') between the three populations. **c, d,** The allele frequency spectra of variants in different portions of the Venn diagrams depicted in **a** and

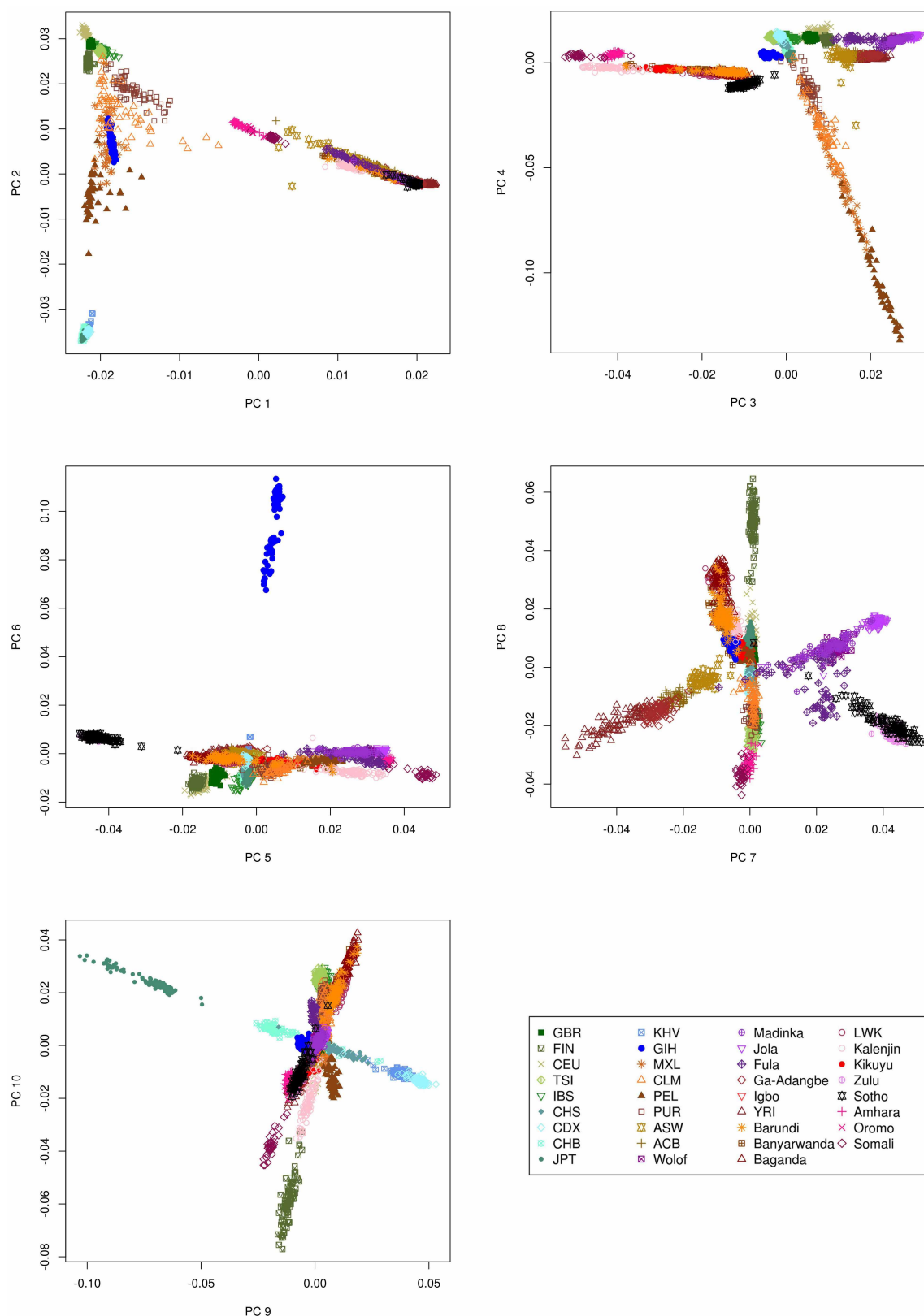
**b,** respectively. There appear to be a large proportion of unshared (private) variants in each population: between 10% and 23% of the total number of variants in a given population. The proportion of novel variants was high, with Ethiopia showing the greatest proportion of novel variation. Most of the novel variation appears to be unshared and rare.



**Extended Data Figure 2 | The first ten principal components for the African data set.** PC1 shows a cline among several African populations, most likely to represent Eurasian gene flow ( $n = 1,481$ ). PC2 shows a clear separation

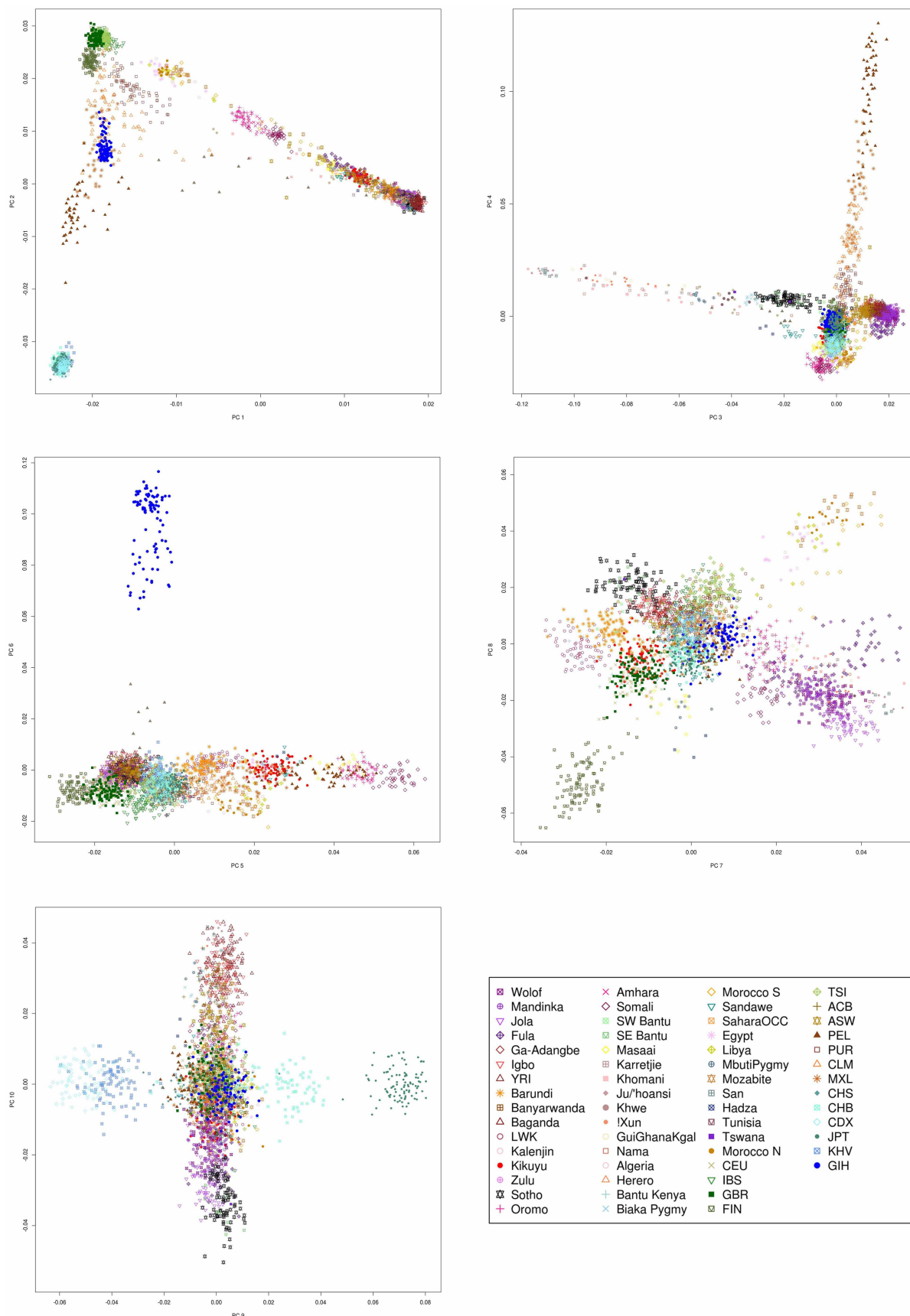
between West and South/East Africa. Subsequent PCs show more detailed structure between, and within African populations.





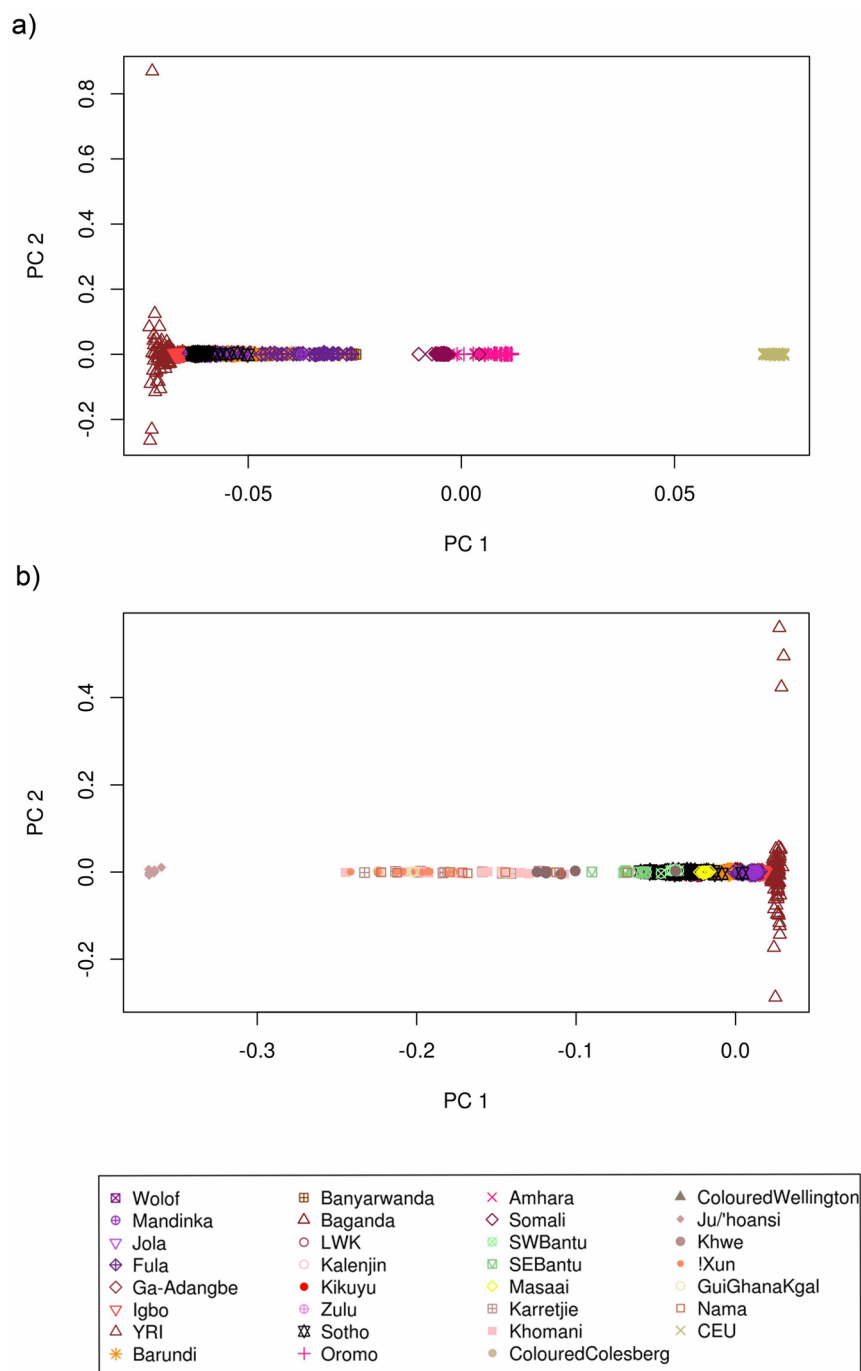
**Extended Data Figure 3 | The first ten principal components for the global data set, including populations from the 1000 Genomes Project.** PC1 shows a cline among several African populations extending towards European populations, most likely to represent non-SSA gene flow ( $n = 2,864$ ). PC2 shows a clear separation between European and Asian populations. Subsequent PCs show more detailed structure between populations globally, and within African populations. GBR, British in England and Scotland; ACB, African Caribbeans in Barbados; ASW, Americans of African ancestry in southwestern

USA; CDX, Chinese Dai in Xishuangbanna, China; CEU, Utah residents with Northern and Western Han European ancestry; CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese; CLM, Colombians from Medellin, Colombia; FIN, Finnish in Finland; GIH, Gujarati Indian from Houston, Texas, USA; IBS, Iberian population in Spain; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam; MXL, Mexican ancestry from Los Angeles, USA; PEL, Peruvians from Lima, Peru; PUR, Puerto Ricans from Puerto Rico, and TSI, Toscani in Italy.



**Extended Data Figure 4 |** The first ten principal components for the global extended data set, including populations from the 1000 Genomes Project, Human Genome Diversity Project, North African and Khoe-San population groups. PC1 shows a cline among several African populations

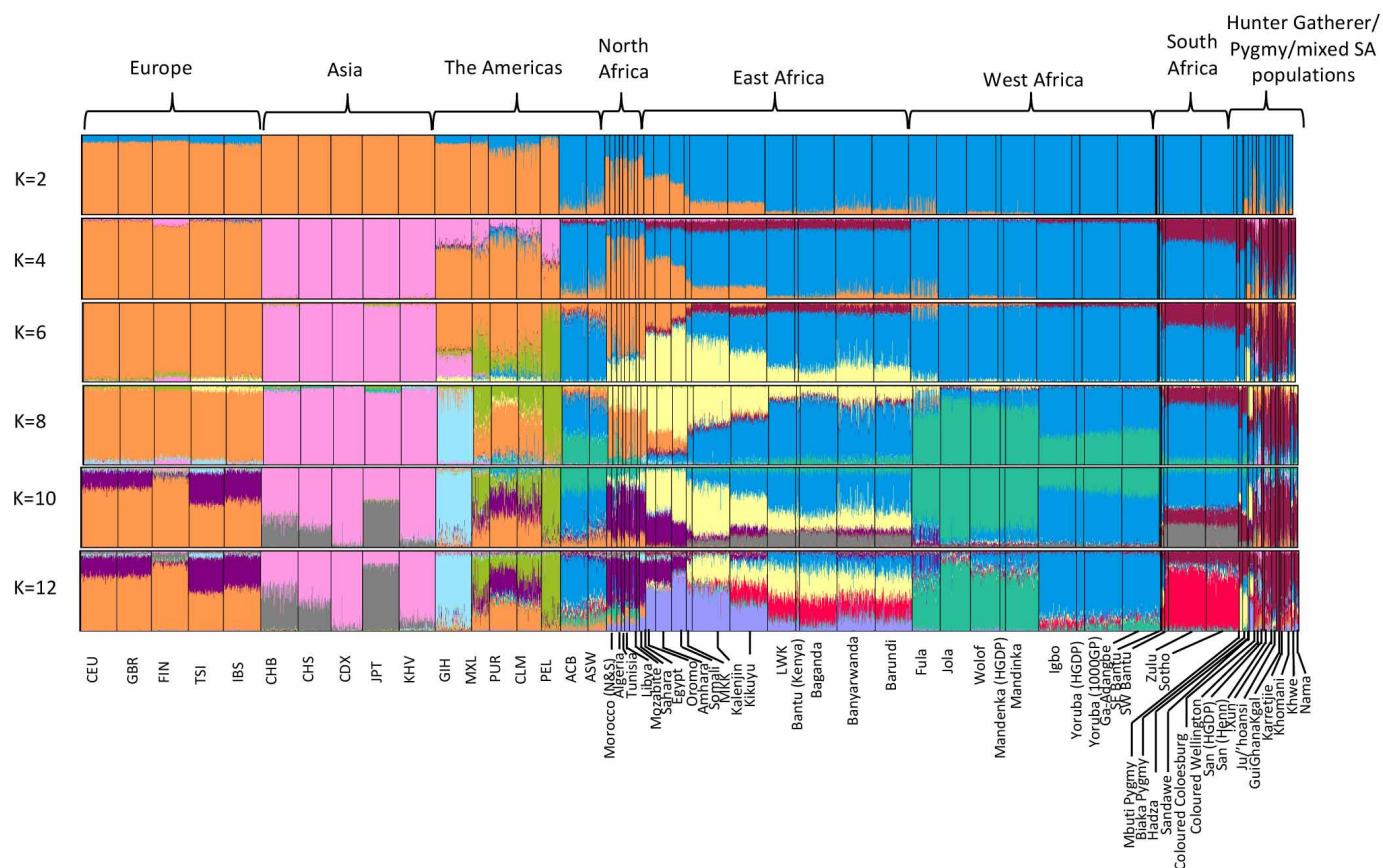
extending towards European populations, most likely to represent non-SSA gene flow ( $n = 3,202$ ). PC2 shows a clear separation between European and Asian populations. Subsequent principal components show more detailed structure between populations globally, and within African populations.



**Extended Data Figure 5 | Projection of principal components to assess admixture among African populations.** **a**, The projection of principal components calculated on YRI and CEU from the 1000 Genomes Project onto the African populations. The AGVP populations are seen to fall on a cline between YRI and CEU, with Ethiopian populations closest to CEU. This is suggestive of Eurasian ancestry among these populations. **b**, The projection of

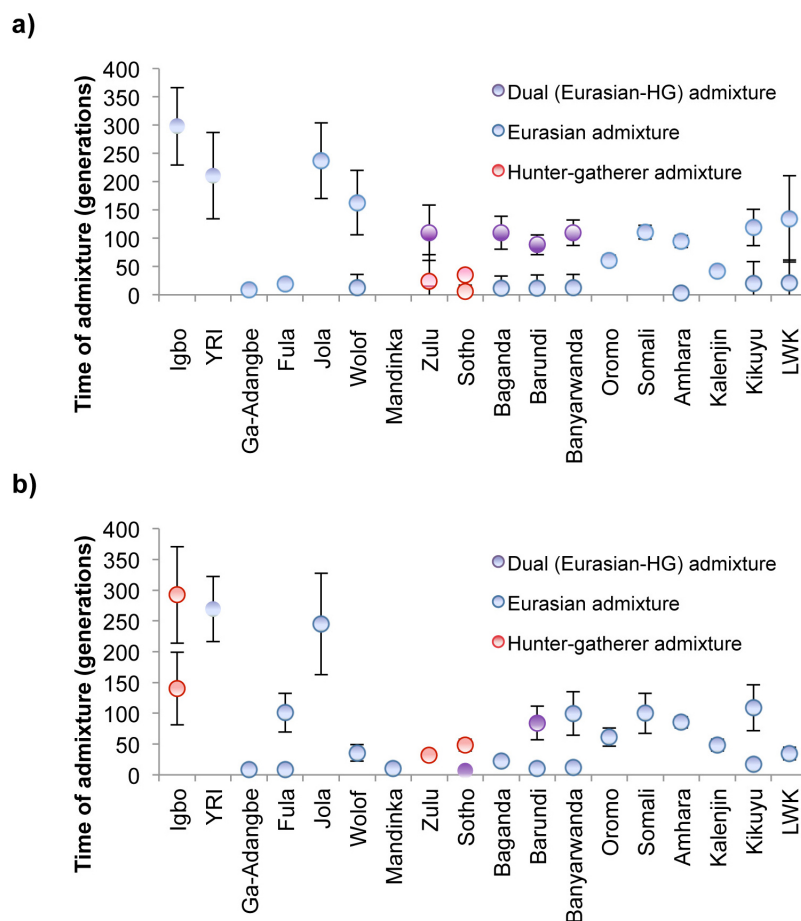
principal components calculated on YRI and Ju/'hoansi onto the AGVP and other Khoe-San populations. The AGVP and Khoe-San populations are seen to fall on a cline between YRI and Ju/'hoansi, with Zulu and Sotho leading the cline among the AGVP populations. This is suggestive of HG gene flow among these populations.





**Extended Data Figure 6 | ADMIXTURE clustering analysis for AGVP samples combined with the 1000 Genomes Project, Human Genome Diversity Project, North African and Khoe-San samples. Cluster  $K = 2$**

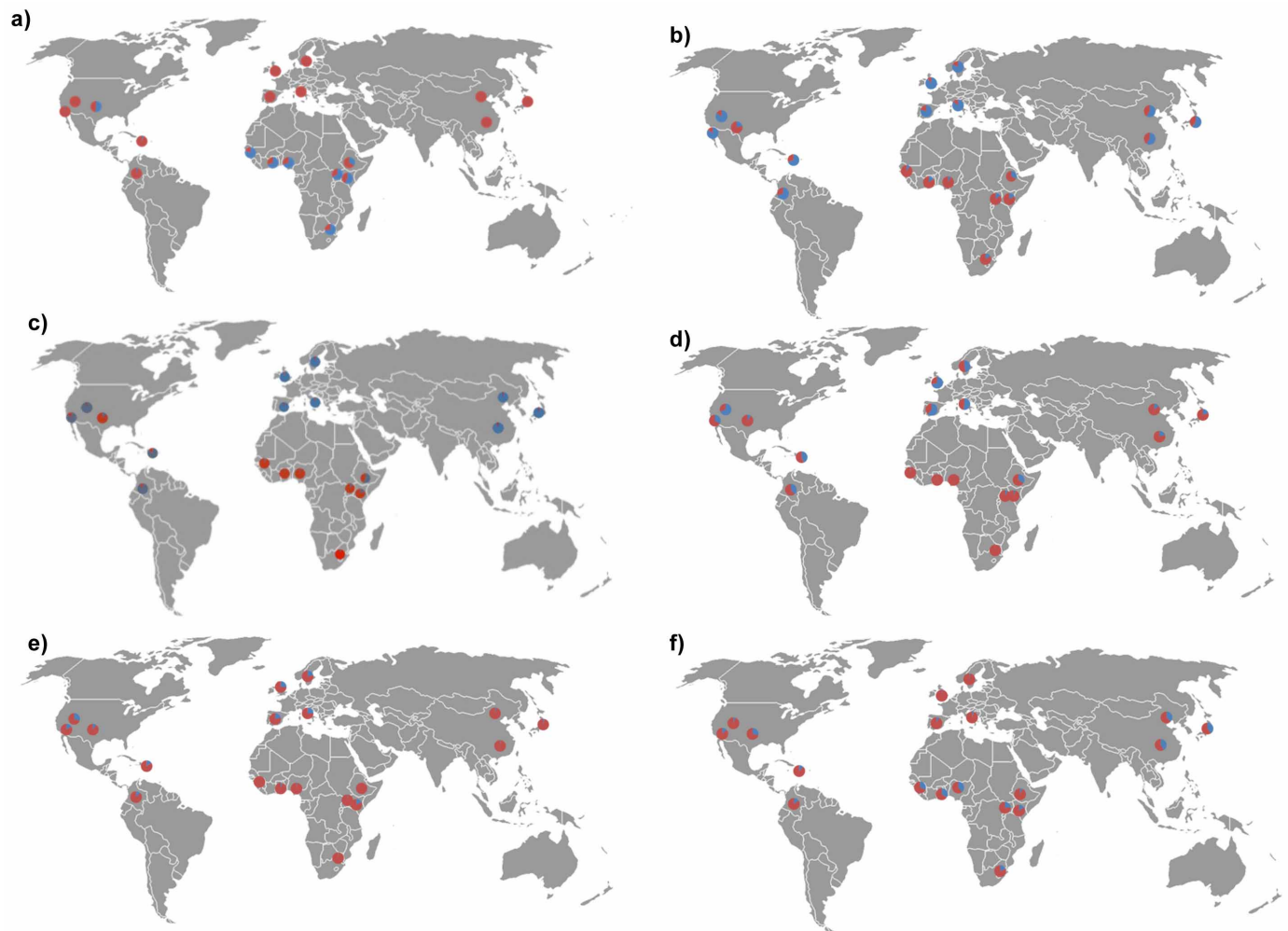
shows separation of European and African ancestry, with delineation of Asian and Khoe-San ancestry in cluster  $K = 4$ . Subsequent clusters show separation of East, West, North and South African ancestral components  $n = 3,202$ .



**Extended Data Figure 7 | Dating and source of admixture in the AGVP.**

**a.** The time and most likely sources of admixture with means and 95% confidence intervals for different AGVP populations estimated with MALDER (see Supplementary Note 5). Circular markers with a line drawn around them

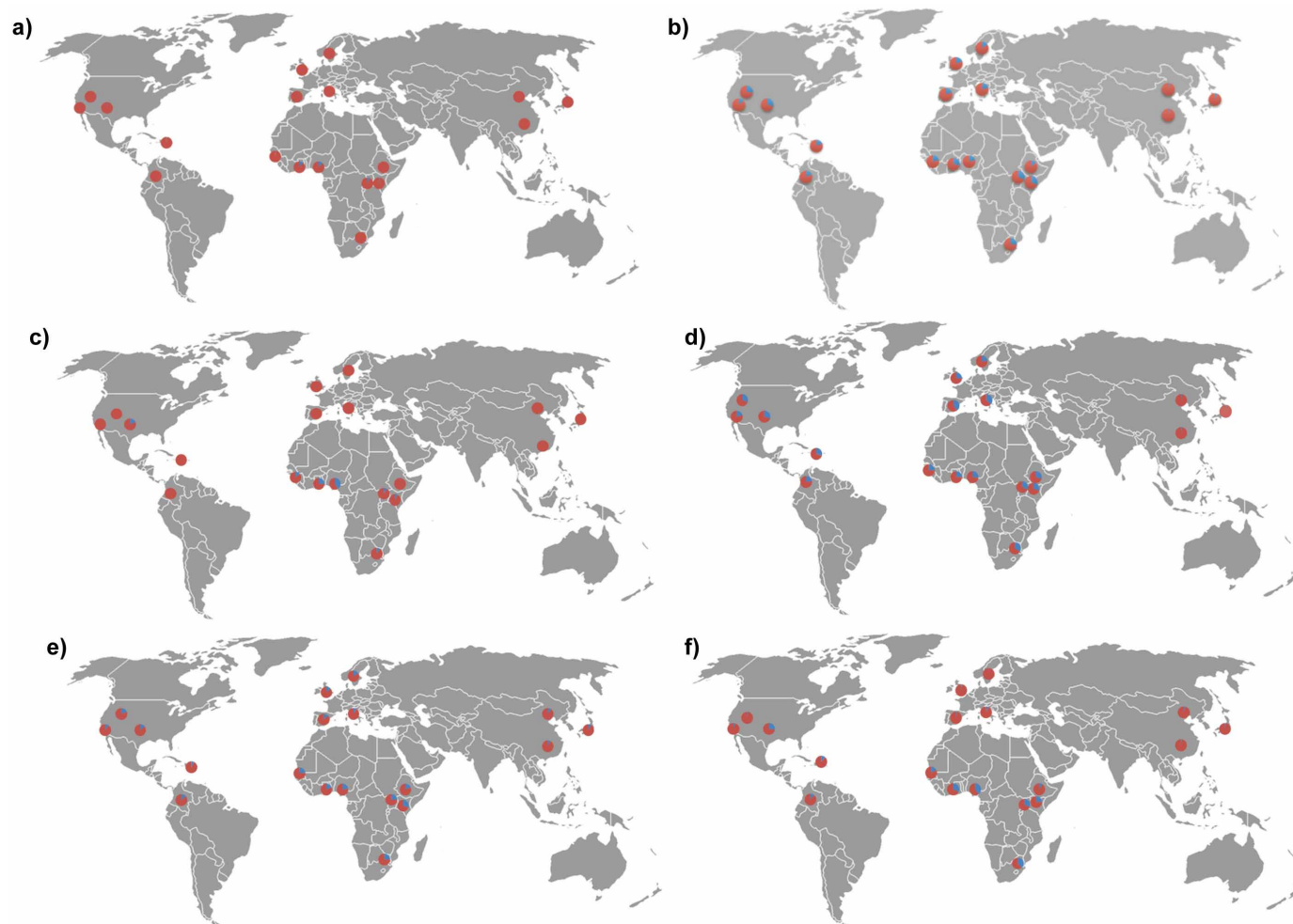
represent high-probability events, while those with no line around them represent low-probability events. **b.** The time and most likely sources of admixture estimated with MALDER for the same populations using high-quality imputed data to improve resolution.



**Extended Data Figure 8 | Loci with marked allelic differentiation either globally or within Africa.** The derived and ancestral alleles are depicted in blue and red, respectively, for all loci. **a**, The global distribution of the non-synonymous variant rs17047661 at the *CRI* locus implicated in malaria severity. This locus was noted to be among the most differentiated sites (in the top 0.1%) between Europe and Africa. **b**, The global distribution of the rs10216063 SNP at the *AQP2* locus. The derived allele appears to be the major allele among European populations in contrast to African populations. **c**, The allele frequency distribution of rs10924081 at the *ATP1A1* locus. Marked

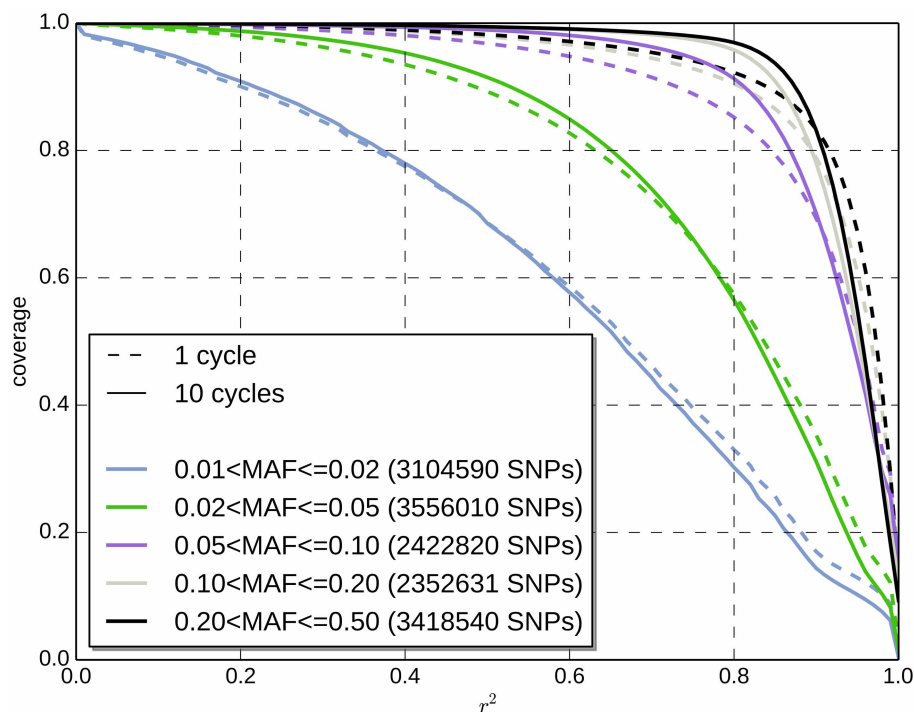
differentiation is observed globally, with the derived allele noted to be the major allele among European populations. **d**, The global distribution of the risk allele for the SNP rs1378940 in the *CSK* locus associated with hypertension. This locus was found to be within the top 0.1% of differentiated loci within Africa, and within the top 1% of differentiated loci globally. **e**, The allele frequency distribution of the rs3213419 SNP at the *HP* locus. **f**, The allele frequency distribution of the rs7313726 SNP at the *CD163* locus. The *HP* and *CD163* are among the top 0.1% of differentiated sites between malaria endemic and non-endemic regions in Africa.





**Extended Data Figure 9 | The global distribution of biologically relevant loci used for simulation of traits to examine reproducibility of signals across AGVP populations.** **a**, The frequency of the sickle-cell variant (rs334) in different regions globally. The blue portion of each pie chart represents the

frequency of the causal allele A. **b**, The distribution of the *SORT1* causal SNP rs12740374, with the derived allele T depicted in blue. **c–f**, The distributions of the *APOL1* variant rs73885319, *TCF7L2* variant rs7903146, the *APOE* variant rs429358 and the *PRDM9* variant rs6889665, respectively.



**Extended Data Figure 10 | The coverage obtained across the genome for variants at different allele frequencies for a hypothetical African genotype array with one million tagging variants.** Different allele frequency bins are depicted in different colours. The lines show the coverage that can be achieved by imputation at different  $r^2$  thresholds. Coverage, here, is defined as the proportion of variants within an allele frequency captured above a pre-defined  $r^2$  threshold (along the x axis) after imputation. The solid lines represent the

coverage obtained with one million variants selected using the hybrid tagging and imputation approach, while the broken lines represent the coverage obtained by using a simple pairwise tagging approach to capture one million tagging variants. The hybrid method improves the coverage obtained, particularly for common variation. Coverage for common variants (>5%) appears to be high at an  $r^2$  threshold of 0.8 and above, with >80% of these variants accurately imputed.

# Internal models direct dragonfly interception steering

Matteo Mischiati<sup>1\*</sup>, Huai-Ti Lin<sup>1\*</sup>, Paul Herold<sup>1</sup>, Elliot Imler<sup>2</sup>, Robert Olberg<sup>3</sup> & Anthony Leonardo<sup>1</sup>

**Sensorimotor control in vertebrates relies on internal models. When extending an arm to reach for an object, the brain uses predictive models of both limb dynamics and target properties. Whether invertebrates use such models remains unclear. Here we examine to what extent prey interception by dragonflies (*Plathemis lydia*), a behaviour analogous to targeted reaching, requires internal models. By simultaneously tracking the position and orientation of a dragonfly's head and body during flight, we provide evidence that interception steering is driven by forward and inverse models of dragonfly body dynamics and by models of prey motion. Predictive rotations of the dragonfly's head continuously track the prey's angular position. The head-body angles established by prey tracking appear to guide systematic rotations of the dragonfly's body to align it with the prey's flight path. Model-driven control thus underlies the bulk of interception steering manoeuvres, while vision is used for reactions to unexpected prey movements. These findings illuminate the computational sophistication with which insects construct behaviour.**

Prediction and planning, essential to the high-performance control of behaviour, require internal models<sup>1</sup>. Decades of work in humans and non-human primates have provided evidence for three types of internal models that are fundamental to sensorimotor control: physical models to predict properties of the world<sup>2,3</sup>; inverse models to generate the motor commands needed to attain desired sensory states<sup>4</sup>; and forward models to predict the sensory consequences of self-movement<sup>5,6</sup>. Without such model-driven control, sensory latencies would leave us grasping at things that had already moved and motor complexity would leave us fumbling with our limbs in the wrong configuration. However, while even basic reaching actions in vertebrates<sup>3</sup> require predictive control, it is unknown whether insects rely on internal models to guide actions<sup>7,8</sup>.

Prey interception is a behaviour that is common to many insects, and is functionally similar to reaching. Because prey make unexpected turns, the predator's interception course cannot be pre-planned, and the control of steering has been viewed as purely reactive. Dragonflies<sup>9–11</sup> in particular are thought to use a classic interception strategy known as parallel navigation<sup>12,13</sup> (Fig. 1a). In parallel navigation, the pursuer steers so as to steadily reduce the length of the range vector from pursuer to prey while holding its direction constant. When prey angular velocity rotates the range vector, it elicits a reaction from the pursuer to stabilize the range vector direction (Fig. 1b). The appeal of parallel navigation is that it is time optimal<sup>14</sup>, computationally simple<sup>13</sup>, and requires no internal models of the body or world. However, the movements forming the interception trajectory have a speed, accuracy and complexity<sup>15</sup> rivaling those of vertebrate reaching tasks<sup>5,16</sup>. This suggests that either the dragonfly's reflexes and manoeuvrability are exquisitely honed, or that steering is not purely reactive but is instead guided by model-driven control.

We thus sought to determine whether dragonfly prey interception is dependent on internal models. We constructed a flight arena in which we could elicit interception flights to computer-controlled prey (Extended Data Fig. 1), and used microscopic retroreflective markers attached to the dragonfly, along with motion-capture techniques, to determine what

the dragonfly saw and how its body moved during flight. The timing and structure of these head and body movements revealed that they were generated by predictions from forward, inverse and target models, combined with visual reactions to unexpected prey manoeuvring. These internal models provide the computational power to simultaneously estimate prey motion and determine the motor commands needed to compose an interception trajectory satisfying biomechanical constraints.

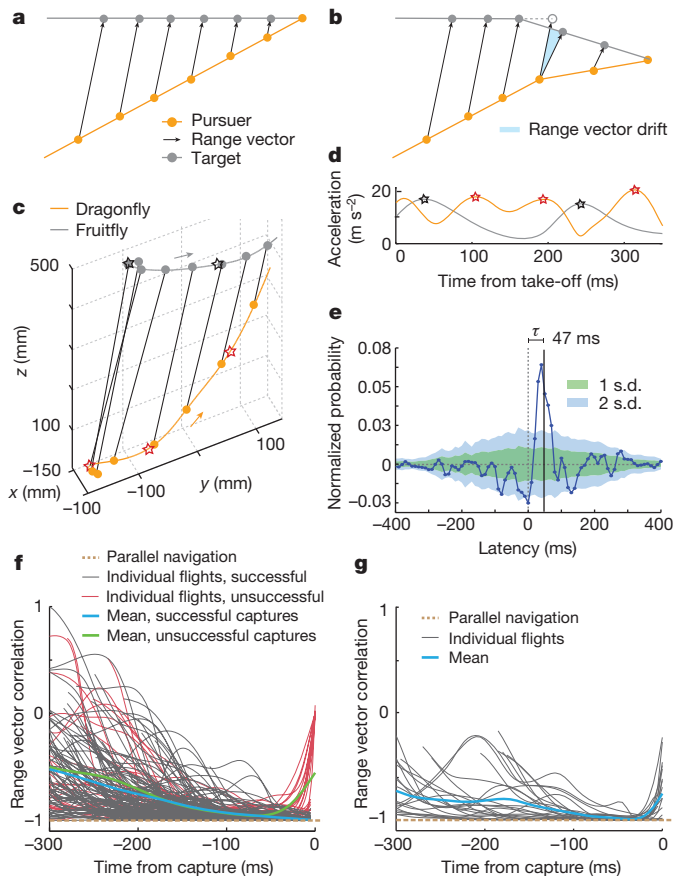
## Reactive steering to prey angular velocity

We began our investigation of model-driven control in insect behaviour by evaluating the accuracy with which simpler reactive strategies explain dragonfly interception steering. The basis for the parallel navigation hypothesis in dragonflies<sup>9–11</sup> is the observation that their flight paths hold the range vector to the prey at roughly constant angle (Fig. 1c). We assessed the fit of parallel navigation by examining the range vector and its derivative—parallel navigation requires them to be exactly anti-correlated, pointing in opposite directions, so that their sum yields a smaller range vector with the same direction. In the absence of prey manoeuvring, a pursuer implementing parallel navigation will quickly drive the range vector correlation to  $-1$ . We calculated the range vector correlation for two data sets: interception flights to fruitflies (Supplementary Videos 1 and 2), and interception flights to computer-controlled constant-velocity artificial prey (2 mm bead, speed  $0.2\text{--}1.5\text{ m s}^{-1}$ ; Supplementary Videos 3 and 4, Methods). The former test whether parallel navigation can be attained with manoeuvring prey that perturb the range vector direction, whereas the latter test how much faster parallel navigation can be attained with non-maneuvring prey that never disturb the range vector. In both cases we found that parallel navigation (that is, correlations close to  $-1$ ) occurred reliably only near capture (Fig. 1f, g). Notably, the mean time evolution of the range vector correlation for constant-speed prey was strikingly similar to that of real prey. Furthermore, the geometry of many interceptions approached parallel navigation and then deviated from it (Fig. 1g). As constant-speed prey do not manoeuvre, these deviations must have arisen from

<sup>1</sup>Janelia Research Campus, Howard Hughes Medical Institute; 19700 Helix Drive, Ashburn, Virginia 20147, USA. <sup>2</sup>University of Arizona, Department of Neuroscience, 1040 E. 4th Street, Tucson, Arizona 85721, USA. <sup>3</sup>Union College, 807 Union Street, Schenectady, New York 12308, USA.

\*These authors contributed equally to this work.





non-parallel-navigation manoeuvring by the dragonfly, and therefore rule out parallel navigation as the steering strategy underlying dragonfly interception.

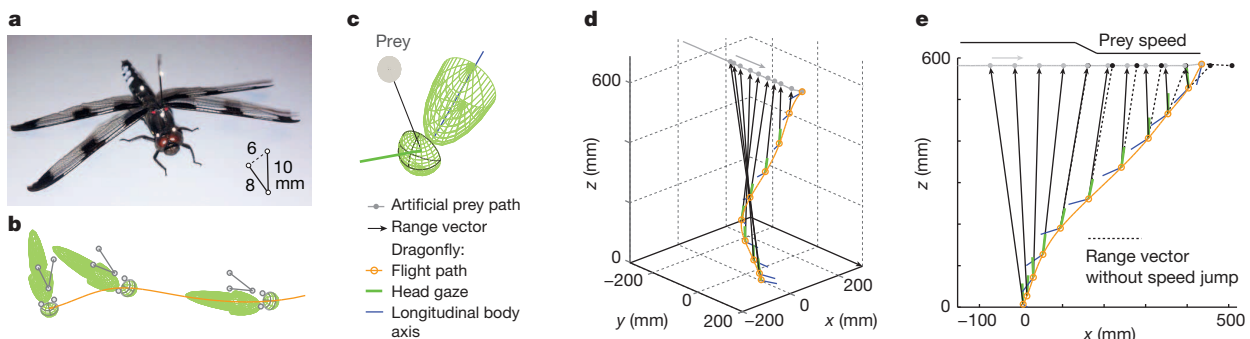
We excluded other reactive steering strategies based on prey angular velocity by evaluating the general prediction of such models that every significant prey manoeuvre is rapidly met by a corrective manoeuvre from the pursuer (Extended Data Fig. 2a). To look for such a temporal correlation between prey and pursuer steering, we identified the timing of the largest dragonfly and prey steering events as peaks in their acceleration (Fig. 1d and Methods). We then compared the distribution of time lags between pairs of prey and dragonfly steering events to a control distribution that assumed steering events were random and

**Figure 1 | Parallel navigation and reactive control are insufficient to describe interception steering.** **a**, The principle of parallel navigation is to hold the range vector at a fixed angle while reducing its length (schematic). **b**, Prey manoeuvring produces drift in the range-vector angle, which will be nullified by a pursuer using parallel navigation thereby establishing a new interception course (schematic). **c**, Three-dimensional (3D) flight path of dragonfly pursuing a fruitfly (50 ms steps). Stars indicate large steering events. **d**, Acceleration magnitude for the data shown in **c**. Stars indicate large steering events (see Methods). **e**, Distribution of latencies between prey ( $t = 0$ ) and dragonfly steering events (155 prey events, 18 dragonflies), normalized to occurrence rates expected for chance levels. Dragonfly steering events exceeding the 2 s.d. margins occurred  $47 \pm 13$  ms after prey steering (range 30–70 ms), accounting for 31% of prey events. **f**, Range vector correlation for flights against real prey, for 110 successful captures (correlation =  $-0.99 \pm 0.02$  at capture) and 30 unsuccessful flights (correlation =  $-0.68 \pm 0.31$  at closest point to prey). **g**, Range vector correlation for constant-speed artificial prey (26 flights, 8 dragonflies, correlation =  $-0.98 \pm 0.04$  just before capture).

independent (see Methods). The difference between these distributions (Fig. 1e) revealed a significant excess of events with a mean lag of  $47 \pm 13$  ms ( $\pm$  s.d., used throughout), providing an estimate of the reactive visuomotor latency through the dragonfly nervous system. However, approximately 70% of the prey steering events were not followed by a detectable dragonfly response, suggesting that in these cases the dragonfly response was either disproportionately small or occurred with a variable latency. Likewise, approximately 75% of the dragonfly steering events were not associated with a prey steering event. These observations directly contradict purely reactive control strategies, which predict a one-to-one mapping between the timing of prey and pursuer steering.

### Body orientation directs steering

We next investigated the purpose of the dragonfly steering events that were not directly related to prey angular velocity. While steering can also be based on other aspects of prey motion, such as angular position<sup>17,18</sup>, it need not be based exclusively on prey motion. For example, the dragonfly's body orientation might influence its manoeuvrability and thereby the structure of its flight path. Likewise, the orientation of the high-acuity zone of its eye might affect the detection and response to prey manoeuvres<sup>19</sup>. Accommodating these constraints would require that the dragonfly relies on internal models of its head and body, which could be used reactively or predictively. To determine whether such models guide steering, we designed microscopic retroreflective markers that could be attached to a dragonfly's head and body and tracked during flight. This enabled us to reconstruct the time-varying three-dimensional head and body orientation of a foraging dragonfly, along with its position and that of its prey (see Methods, Fig. 2a–d and Extended Data Fig. 3).



**Figure 2 | Free-flight measurement of head and body orientation during prey interception.** **a**, Dragonfly (*Plathemis lydia*) with 750  $\mu$ m retroreflective markers on head (just above the red-tinted fovea of the eye) and 1,000  $\mu$ m markers on body. **b**, 3D-in-flight marker data, reconstructed from an 18-camera array (200 frames-per-second (fps), 150  $\mu$ m root-mean-square (RMS) tracking accuracy in a  $\sim 2$  m<sup>3</sup> volume; see Extended Data Figs 1 and 3). **c**, Dragonfly

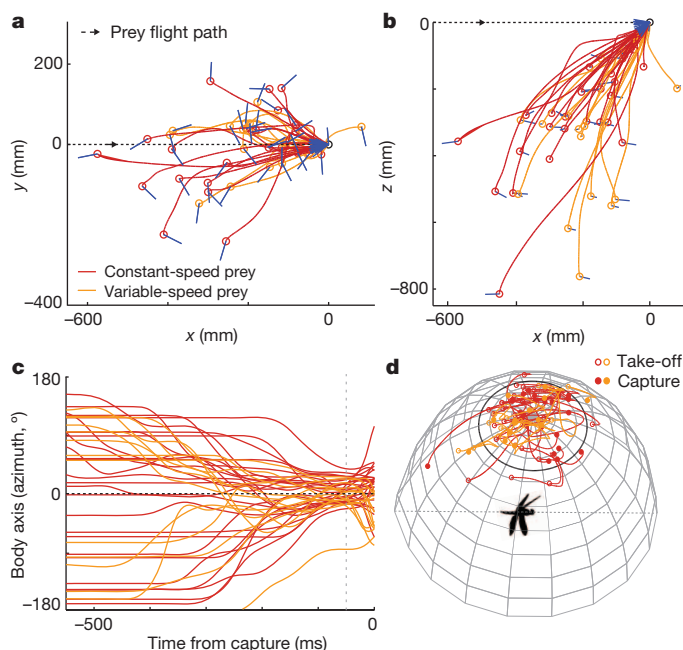
body axis, direction of gaze (foveation) and direction to prey (range vector). Black contour indicates foveal region of the eye. **d**, Interception flight path with head and body orientations (50 ms steps); prey remains foveated throughout  $1,000^\circ \text{ s}^{-1}$  turn. **e**, Interception flight to variable-speed artificial prey. Black circles indicate the trajectory of the prey had the speed jump not occurred.

To evaluate whether dragonfly body orientation constrained the flight path, we analysed how this orientation evolved over the course of the flight. Constant-speed prey allowed us to identify changes in body orientation that were unrelated to prey manoeuvres. We found that systematic rotations of the dragonfly's body often made the flight path significantly different from the straight-line path that would have minimized the distance to the same interception point (Fig. 3). During interception flights, dragonflies maintained their azimuthal bearing aligned to their body axis ( $-5^\circ \pm 29^\circ$ ). As each dragonfly perched in a random orientation relative to the prey's flight path, to fly towards the interception point typically required a turn. These turns aligned the dragonfly's body axis to the prey's direction of motion in azimuth, while stabilizing the elevation of the body axis at about  $30^\circ$  (Fig. 3a–c). Such a stereotyped alignment across flights suggests a specific body orientation is preferred for grasping the prey. As these manoeuvres occurred, the dragonfly maintained the prey's angular position to within  $25^\circ$  of directly overhead (Fig. 3d and Supplementary Video 5). The interception trajectory thus positioned and aligned the dragonfly directly below the prey and thereby reduced the capture problem to one of closing the vertical gap with the body in a constrained orientation.

When dragonflies were confronted with prey that made a 20–80% speed jump mid-flight, their steering manoeuvres were virtually indistinguishable from those to constant-speed prey. This suggests that visual responses to prey manoeuvres were small corrections integrated into the ongoing steering program focused on body alignment and prey angular position (Figs 2e and 3 and Supplementary Video 6).

### Head orientation tracks prey position

Dragonfly body rotations induce substantial apparent prey motion that complicates detection of the prey's angular position. This problem is



**Figure 3 | Body dynamics constrain the interception steering strategy.** **a, b,** Top (xy) and side view (xz) of flight trajectories to artificial prey, including dragonfly body axis orientation (40 flights, 26 constant-speed, 14 variable-speed trials, 9 dragonflies;  $x = y = z = 0$  is the trial alignment point, 50 ms before capture). The body axis makes a  $5 \pm 17^\circ$  angle in azimuth and a  $22 \pm 18^\circ$  angle in elevation with the constant-speed prey flight path just prior to capture ( $-6 \pm 27^\circ$  azimuth and  $36 \pm 16^\circ$  elevation, variable-speed prey). Note that at take-off the body starts in a random orientation. **c,** Time evolution of the azimuthal angle between the dragonfly's body axis and prey's flight path, for the population data shown in **a**. Vertical grey line is the alignment point used in **a, b**. **d,** Spherical projection of prey flight paths relative to dragonfly position over the interception flight. North pole represents the zenith; black line the  $25^\circ$  contour.

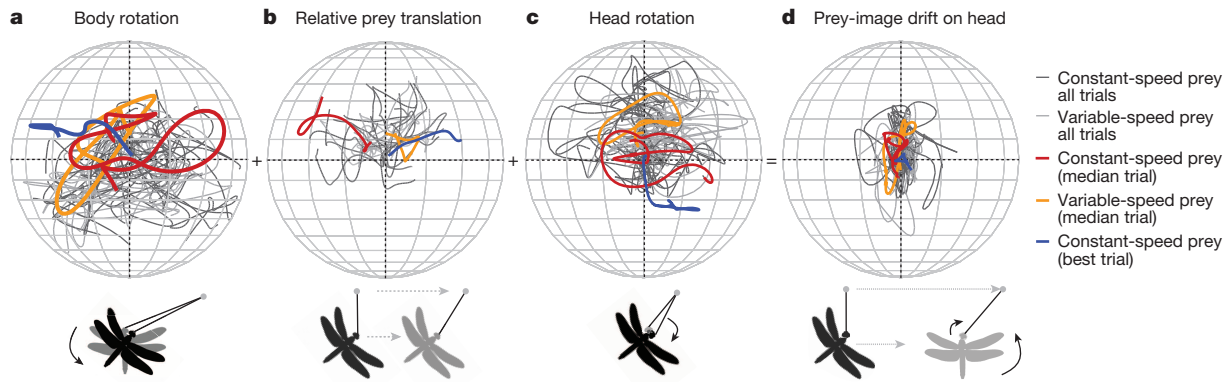
amplified because the distance from dragonfly to prey scales the prey-image drift from prey manoeuvres, but has no influence on that from dragonfly body rotations. Thus, if the dragonfly rotates its body at  $1,000^\circ \text{ s}^{-1}$  (for example, Fig. 2d and Supplementary Video 5) and the prey makes an abrupt turn at  $1 \text{ m s}^{-1}$ , the angular shift in prey position from dragonfly motion can be  $10\times$  larger than that from the prey itself. Discriminating prey-image drift that arises from prey motion rather than dragonfly motion is the cost of not using a steering strategy, such as parallel navigation, that holds prey position fixed relative to a rigidly aligned head and body. This problem can be solved through signal processing within the nervous system to cancel apparent motion, or through movements of the head to nullify such motion before detection by neurons. We sought to determine which solution was used and to what extent it relied on prediction versus reaction.

To determine how the dragonfly solves the drift discrimination problem, we used our motion-capture system to estimate prey-image drift on the dragonfly's eye during interception flights. We first measured the orientation of the dragonfly's head and then derived the angular position of the prey image on the dragonfly's compound eye (which is rigidly fixed to the head). Separately, we quantified the three underlying sources of prey-image drift: rotation of the dragonfly's body; relative translation of the dragonfly and prey; and rotation of the dragonfly's head relative to the body (Fig. 4 and Extended Data Fig. 4). We found that the prey image was held steady on the dorsal surface of the head (typically within  $4^\circ$  azimuth and  $7^\circ$  elevation) in the approximate location of the high-resolution optical fovea (Figs 2a and 4d)<sup>10,19</sup>. We refer to the process of holding the image stable on the head as 'foveation'. The prey-image drift on the head was on average less than 50% that from body rotations, indicating significant cancellation. Indeed, head rotations were a linear function of the angular velocity induced by body rotations and relative translation, with a slope close to  $-1$  over a range of  $\pm 1,000^\circ \text{ s}^{-1}$  (Fig. 5a, b, d, e and Supplementary Videos 7 and 8). These data confirm that apparent motion disturbances are nullified through head movements before they enter the nervous system.

### Predictive foveation

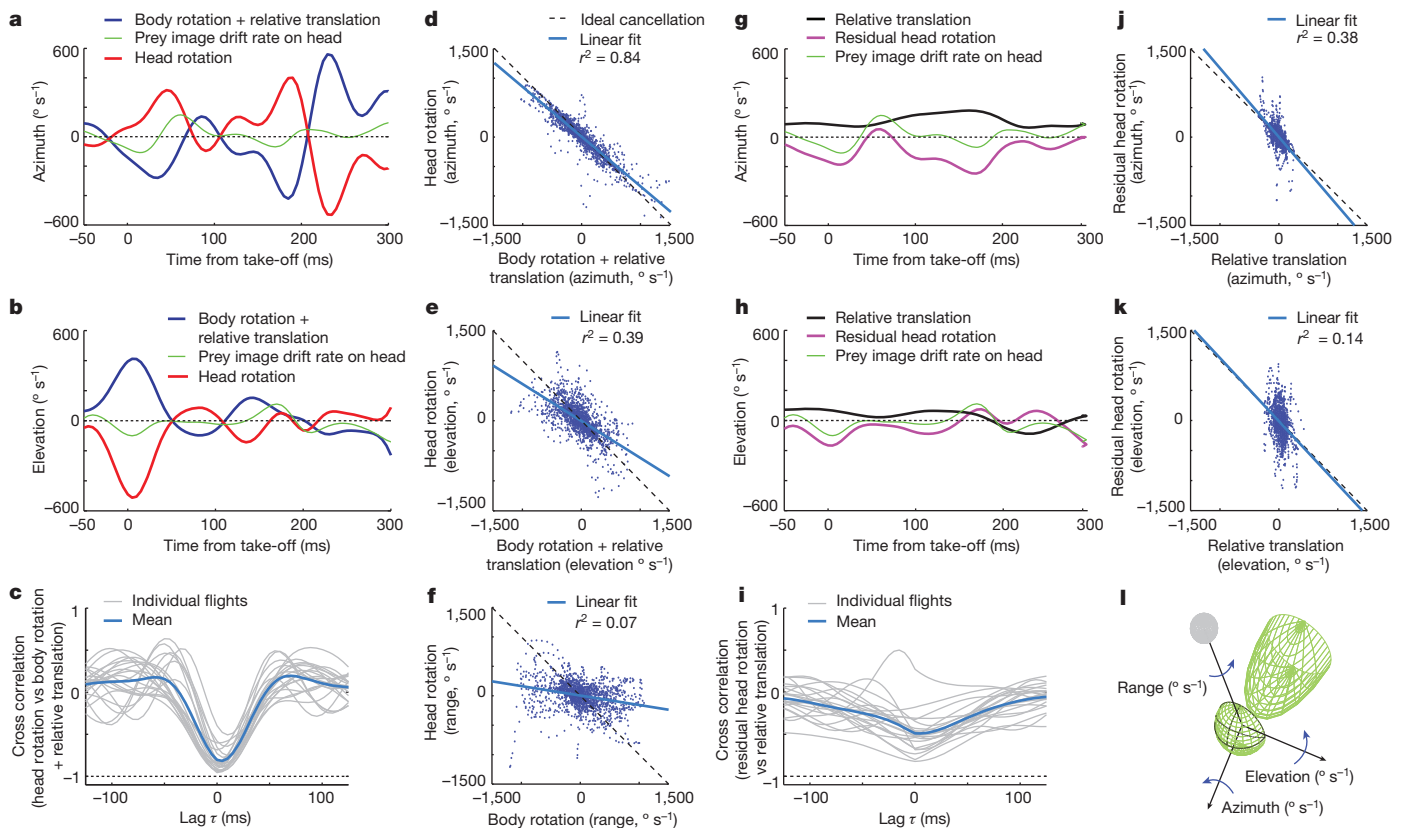
The compensatory head rotations could be driven reactively, based on sensory feedback from prey-image drift, or predictively, based on internal models of drift. To discriminate between these two possibilities, we examined the delay between the summed foveation disturbances (Fig. 4a, b) and the corrective head rotations (Fig. 4c). Insect muscle contractions require approximately 5 ms to produce force<sup>20</sup>, and any visually driven reactive movements of the head should display an even larger sensorimotor lag (for example, 47 ms from vision to wing, Fig. 1e). In contrast, if the dragonfly predicted foveal drift, the head could be moved with zero delay for optimal cancellation, especially in the constant-speed prey condition in which there is no unexpected manoeuvring. We found that head movements occurred with an average lag of  $4 \pm 4$  ms after the disturbance they cancelled (Fig. 5c). Such an exceedingly brief delay is strong evidence for the head being controlled predictively rather than reactively. This prediction of prey-image motion was specific to the goal of foveation. Head rotations nullified body rotations that created prey-image drift in azimuth and elevation, but not body rotations about the range vector axis, which has no influence on foveation (Fig. 5l; compare Fig. 5d, e vs Fig. 5f). The restriction of cancellation to the subspace of behavioural axes relevant for foveation is consistent with well-known optimality principles for sensorimotor control<sup>21,22</sup>.

To predict prey-image drift accurately, the dragonfly should model both of its individual sources: drift from body rotations (Fig. 4a) and drift from relative prey translation (Fig. 4b). Because the sum of these disturbances is dominated by drift induced by body rotations ( $\pm 1,000^\circ \text{ s}^{-1}$  range, Fig. 5d, e versus  $\pm 200^\circ \text{ s}^{-1}$  range for relative prey translation, Fig. 5j, k), the compensatory head rotations clearly contained a negative copy of the body rotation signal. To assess the strength of cancellation of translational disturbances, we calculated the residual head rotation signal; that is, the sum of the head rotation and body rotation prey-image



**Figure 4 | Decomposition of prey-image drift into its sources reveals cancellation.** Dragonfly head-centred spherical projections of artificial prey trajectories during interception (37 flights, 21 constant-speed prey, 16 variable-speed prey, 5 dragonflies). The intersection of the dashed black lines shows the anatomical mid-line, in longitude, and the  $38^\circ$  latitude (horizontal); this corresponds approximately to the centre of the optical fovea (see Fig. 2c). Drift magnitude is reported in root-mean-square azimuth ( $\text{RMS}_{\text{azim}}$ ) and elevation ( $\text{RMS}_{\text{elev}}$ ) for constant-speed and variable-speed prey across trials. **a**, Prey-image drift from rotations of the dragonfly's body:  $17.1^\circ \text{RMS}_{\text{azim}}$  and

$12.8^\circ \text{RMS}_{\text{elev}}$ , for constant-speed prey ( $17.3^\circ \text{RMS}_{\text{azim}}$  and  $12.6^\circ \text{RMS}_{\text{elev}}$ , for variable-speed prey). **b**, Prey-image drift from relative prey translation:  $8.4^\circ \text{RMS}_{\text{azim}}$  and  $7.4^\circ \text{RMS}_{\text{elev}}$ , for constant-speed prey ( $5.2^\circ \text{RMS}_{\text{azim}}$  and  $5.7^\circ \text{RMS}_{\text{elev}}$ , for variable-speed prey). **c**, Prey-image drift from head rotations relative to the body:  $11.7^\circ \text{RMS}_{\text{azim}}$  and  $9.8^\circ \text{RMS}_{\text{elev}}$ , for constant-speed prey ( $14.6^\circ \text{RMS}_{\text{azim}}$  and  $8.0^\circ \text{RMS}_{\text{elev}}$ , for variable-speed prey). **d**, Prey-image drift on the head and eye, equal to the sum of **a–c**:  $3.8^\circ \text{RMS}_{\text{azim}}$  and  $7.0^\circ \text{RMS}_{\text{elev}}$ , for constant-speed prey ( $3.4^\circ \text{RMS}_{\text{azim}}$  and  $7.8^\circ \text{RMS}_{\text{elev}}$ , for variable-speed prey). See Extended Data Fig. 4.



**Figure 5 | Head movements predictively compensate for disturbances to foveation.** (5 dragonflies, 21 constant-speed artificial prey flights). **a, b**, Time evolution of the prey-image drift rate from summed foveation disturbances (Fig. 4a, b), dragonfly head rotation (Fig. 4c) and actual drift on the head (Fig. 4d) for the blue trial in Fig. 4. **c**, Cross correlation between summed foveation disturbances and head rotations (mean lag across trials,  $4 \pm 4$  ms). **d, e**, Scatter plots of summed foveation disturbance and head rotation drift rate (all trials). Linear fits are compared to the ideal cancellation indicated by a slope of  $-1$ . Azimuthal disturbances are cancelled by 70% (slope  $-0.84 \pm 0.01$ ) versus 30% for elevational ones (slope  $-0.61 \pm 0.03$ ). **f**, Scatter plot of body

rotation versus head rotation around the range vector shows little cancellation (slope  $-0.16 \pm 0.02$ ). **g–k**, Time evolution (**g, h**), cross correlation (**i**) and scatter plots (**j, k**) of the prey-image drift rate from relative prey translation versus residual head rotations. **i**, The mean lag across trials of the residual head rotations is  $-3 \pm 25$  ms. The cancellation of relative prey translation is not as good as that for body rotations. However, slopes of linear fits are still close to the optima of  $-1$  (**j**, slope  $-1.17 \pm 0.05$  azimuth; **k**, slope  $-1.05 \pm 0.08$  elevation). **l**, Axes of rotation producing azimuthal and elevational prey drift on the head, and the range vector axis about which rotations produce no prey drift (see Extended Data Fig. 4).



drift rates. This signal was free of the large body rotation prediction and could therefore be examined for its correlation to the smaller translational drift. We found that the prey-image angular velocity arising from the residual head rotations was on average equal and opposite to that of the translational drift rate (Fig. 5g, h, j, k and Extended Data Fig. 5, and Supplementary Video 9), indicating that the residual head rotations clearly contained either a predictive or a reactive response to translational drift. Residual head rotations occurred on average  $3 \pm 25$  ms before the translational disturbance they were cancelling (Fig. 5i), far earlier than the visual delay expected for reactive steering (Fig. 1e). These observations argue that the compensatory head movements for translation of constant-speed prey must also be predictive. During the interception of variable-speed prey, foveation was still maintained (Fig. 4d), demonstrating that the dragonfly's head rotated reactively when unexpected prey motion occurred, consistent with earlier work<sup>10</sup>. Head rotations thus seamlessly incorporated both predictive and reactive control, and closely approximated an ideal tracking system.

## Discussion

It is a longstanding question<sup>8</sup> as to whether insects use internal models, such as those known in vertebrate sensorimotor tasks<sup>5</sup>. Traditional views have held that invertebrate behaviours, including prey interception<sup>17</sup>, are based solely on reactive control. Even in mammals and birds, interception is often described with no predictive or model-driven components<sup>14,23</sup>. Dragonfly steering in particular has been thought to be implemented by a neural autopilot<sup>11</sup> that uses prey angular velocity to steer the wings reactively and hold the prey image stable on the eye, thereby implementing parallel navigation<sup>10</sup>. Here we have shown that, in striking disagreement with this classic model, the prey-image drift that would drive such an autopilot is minute and poorly correlated with prey angular velocity (Fig. 4d). Moreover, the dragonfly's steering is inconsistent with both parallel navigation and all other reactive control models driven exclusively by prey movement (Fig. 1).

We have found that the interception flight path is substantially influenced by biomechanical constraints. The dragonfly's steering strategy is to align its body and bearing to the prey's direction of motion while remaining directly below the prey and closing the vertical distance to it (Fig. 3 and Supplementary Video 1). Such a body-centric strategy has numerous consequences. Aligning bearing to body axis should improve the dragonfly's speed and manoeuvrability. Aligning them to the prey's flight path elicits orienting turns early in the flight but ultimately positions the dragonfly such that the prey drops directly into its legs, reducing the need for drastic turns in the final moments of capture. Finally, because the dragonfly approaches the prey from below, it falls in the prey's blind spot and is less likely to elicit an escape response.

Central to the interception strategy is prey foveation, the dynamics of which imply direct model-driven control of the head. Predictive head rotations nullify prey-image drift due to both self-motion and constant-speed prey motion with nearly zero lag (Fig. 5 and Supplementary Videos 8 and 9), but only in the moving visual axes that affect foveation. This contrasts with circuits such as the vestibulo-ocular reflex, which reacts to sensory feedback with a delay and counter-rotates the eyes to cancel self-motion in a fixed coordinate system<sup>24</sup>. Our data are consistent with the dragonfly nervous system using an efference copy<sup>16</sup> of wing motor commands to drive a forward model<sup>5,6</sup> that predicts the prey-image drift from self-motion. The efference copy alone<sup>25</sup> is insufficient for cancellation because steering commands do not have a simple relation to either wing motion<sup>26</sup> or prey-image movement. Likewise, at least a simple model of constant-speed prey motion is needed to predict prey translation. An inverse model would then generate the motor commands to produce the exact head rotation that will cancel these expected disturbances.

Prey foveation also imposes indirect model-driven control over body steering. Because foveation is so effective, there is virtually no prey-image motion on the eye for most of the flight (Fig. 4d and Supplementary Video 8). Consequently, prey angular position and velocity cannot be

measured solely from vision. Instead, foveation encodes prey angular position in the dragonfly's head-body angles. Prey angular velocity is probably computed at take-off, and then estimated from memory by incrementing its current value with any new visual drift. The key parameters for steering thus appear to arise primarily from prediction and estimation processes. Neurons that detect prey-image drift<sup>11,27</sup> would be well suited to reactively correct the steering controller during episodes of brief unexpected prey motion, rather than steer the dragonfly continuously. In this view, if the dragonfly's prediction of prey position were perfect, the prey image would be stabilized to a point for the entire flight and vision would be dispensable, at least for the purposes of interception. A control architecture consistent with these results is suggested in Extended Data Fig. 2b.

We have shown that in solving sophisticated motor control problems, insects rely on both predictive and reactive control, combining forward, inverse and target models with visual feedback. Such internal models, while not previously described in insects<sup>8,25</sup>, are analogous to those used in reaching tasks in vertebrates<sup>5</sup>. How such control processes are implemented is a question that cannot be answered solely from behaviour. For example, while foveation implicitly uses both forward and inverse models, it is unclear whether the dragonfly nervous system uses two distinct representations or a single one with the same function<sup>6</sup>. While neural correlates of internal models have been described in primates<sup>4</sup>, their underlying representations are highly distributed and circuit-level understanding remains elusive. Smaller insects, such as *Drosophila*<sup>28</sup>, have not yet been shown to use such control strategies, and their size would make the requisite free-flight measurements challenging. In contrast, dragonflies have stereotyped and accessible neural circuitry, and are large enough to permit free-flight measurements of position, body state and neural data<sup>29</sup>. The combination of these elements is rare, and the dragonfly may thus be a system where internal models can be dissected behaviourally and then mechanistically reconstructed from analysis of the neural circuit dynamics.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 12 May; accepted 6 November 2014.**

**Published online 10 December 2014.**

- Franklin, D. W. & Wolpert, D. M. Computational mechanisms of sensorimotor control. *Neuron* **72**, 425–442 (2011).
- Zago, M. et al. Internal models of target motion: expected dynamics overrides measured kinematics in timing manual interceptions. *J. Neurophysiol.* **91**, 1620–1634 (2004).
- Flanagan, J. R., King, S., Wolpert, D. M. & Johansson, R. S. Sensorimotor prediction and memory in object manipulation. *Can. J. Exp. Psychol.* **55**, 87–95 (2001).
- Kawato, M. Internal models for motor control and trajectory planning. *Curr. Opin. Neurobiol.* **9**, 718–727 (1999).
- Wolpert, D. M., Ghahramani, Z. & Jordan, M. I. An internal model for sensorimotor integration. *Science* **269**, 1880–1882 (1995).
- Mehta, B. & Schaal, S. Forward models in visuomotor control. *J. Neurophysiol.* **88**, 942–953 (2002).
- Huston, S. J. & Jayaraman, V. Studying sensorimotor integration in insects. *Curr. Opin. Neurobiol.* **21**, 527–534 (2011).
- Webb, B. Neural mechanisms for prediction: do insects have forward models? *Trends Neurosci.* **27**, 278–282 (2004).
- Olberg, R. M., Worthington, A. H. & Venator, K. R. Prey pursuit and interception in dragonflies. *J. Comp. Physiol. A* **186**, 155–162 (2000).
- Olberg, R. M., Seaman, R. C., Coats, M. I. & Henry, A. F. Eye movements and target fixation during dragonfly prey-interception flights. *J. Comp. Physiol. A* **193**, 685–693 (2007).
- Gonzalez-Bellido, P. T., Peng, H., Yang, J., Georgopoulos, A. P. & Olberg, R. M. Eight pairs of descending visual neurons in the dragonfly give wing motor centers accurate population vector of prey direction. *Proc. Natl Acad. Sci. USA* **110**, 696–701 (2013).
- Justh, E. W. & Krishnaprasad, P. S. Steering laws for motion camouflage. *Proc. R. Soc. A* **462**, 3629–3643 (2006).
- Shneydor, N. A. *Missile Guidance and Pursuit: Kinematics, Dynamics and Control* (Elsevier, 1998).
- Ghose, K., Horiuchi, T. K., Krishnaprasad, P. S. & Moss, C. F. Echolocating bats use a nearly time-optimal strategy to intercept prey. *PLoS Biol.* **4**, e108 (2006).
- Combes, S. A., Rundle, D. E., Iwasaki, J. M. & Crall, J. D. Linking biomechanics and ecology through predator-prey interactions: flight performance of dragonflies and their prey. *J. Exp. Biol.* **215**, 903–913 (2012).

16. Azim, E., Jiang, J., Alstermark, B. & Jessell, T. M. Skilled reaching relies on a V2a propriospinal internal copy circuit. *Nature* **508**, 357–363 (2014).
17. Collett, T. S. & Land, M. F. How hoverflies compute interception courses. *J. Comp. Physiol. A* **125**, 191–204 (1978).
18. Haselsteiner, A. F., Gilbert, C. & Wang, Z. J. Tiger beetles pursue prey using a proportional control law with a delay of one half-stride. *J. R. Soc. Interface* **11**, 20140216 (2014).
19. Labhart, T. & Nilsson, D. E. The dorsal eye of the dragonfly sympetrum—specializations for prey detection against the blue sky. *J. Comp. Physiol. A* **176**, 437–453 (1995).
20. Tu, M. & Dickinson, M. Modulation of negative work output from a steering muscle of the blowfly *Calliphora vicina*. *J. Exp. Biol.* **192**, 207–224 (1994).
21. Todorov, E. Optimality principles in sensorimotor control. *Nature Neurosci.* **7**, 907–915 (2004).
22. Scholz, J. P. & Schoner, G. The uncontrolled manifold concept: identifying control variables for a functional task. *Exp. Brain Res.* **126**, 289–306 (1999).
23. Tucker, V. A., Tucker, A. E., Akers, K. & Enderson, J. H. Curved flight paths and sideways vision in peregrine falcons (*Falco peregrinus*). *J. Exp. Biol.* **203**, 3755–3763 (2000).
24. Ito, M. Cerebellar control of the vestibulo-ocular reflex—around the flocculus hypothesis. *Annu. Rev. Neurosci.* **5**, 275–296 (1982).
25. Poulet, J. F. & Hedwig, B. A corollary discharge maintains auditory sensitivity during sound production. *Nature* **418**, 872–876 (2002).
26. Wang, Z. J. & Russell, D. Effect of forewing and hindwing interactions on aerodynamic forces and power in hovering dragonfly flight. *Phys. Rev. Lett.* **99**, 148101 (2007).
27. Adelman, T. L., Bialek, W. & Olberg, R. M. The information content of receptive fields. *Neuron* **40**, 823–833 (2003).
28. von Reyn, C. R. *et al.* A spike-timing mechanism for action selection. *Nature Neurosci.* **17**, 962–970 (2014).
29. Thomas, S. J., Harrison, R. R., Leonardo, A. & Reynolds, M. S. A Battery-free multichannel digital neural/EMG telemetry system for flying insects. *IEEE Trans. Biomed. Circ. Sys.* **6**, 424–436 (2012).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank I. Siwanowicz for advice on neck joint anatomy, and J. Melfi for assistance on kinematic computations. J. Osborne and J. Jordan provided assistance with the retroreflector assembly and artificial prey delivery systems. M. Barbic provided guidance on retroreflector mirroring. D. Parks and the Janelia vivarium provided dragonfly husbandry. We are grateful to V. Jayaraman, A. Karpova, W. Denk and J. Wang for discussions and comments on the manuscript. This work was supported by the Howard Hughes Medical Institute. Additional support to R.O. from AFOSR FA9550-10-1-0472.

**Author Contributions** A.L., M.M. and H.-T.L. designed the study. A.L. and R.O. designed the flight arena and Photron system. A.L. and H.-T.L. designed the motion-capture system. E.I. and P.H. collected the Photron data. H.-T.L. instrumented the artificial prey system and collected the motion-capture data. M.M. and H.-T.L. designed the pre-processing algorithms. M.M. analysed the data with input from A.L. A.L. and M.M. wrote the manuscript with input from H.-T.L. and R.O.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.L. ([leonardoa@janelia.hhmi.org](mailto:leonardoa@janelia.hhmi.org)).

## METHODS

**Animal husbandry.** Experiments were conducted with the dragonfly *Plathemis lydia* (males and females, without distinction). Nymphs were collected from ponds near Janelia Research Campus, reared in an indoor aquatic facility, and fed brine shrimp and black worms. After emergence, dragonflies were inserted in a flight arena reproducing naturalistic conditions (see below), where they foraged on free-flying *Drosophila virilis*. Roughly 90% of the dragonflies foraged regularly and steadily gained weight, often reaching weights >400 mg within 2 weeks (from an average of 220 mg at emergence). For kinematics experiments, we selected animals weighing >300 mg that could easily carry the approximately 15 mg mass comprising retroreflectors and frame (see below).

**Flight arena.** We designed an indoor dragonfly flight arena with naturalistic visual cues and controlled light, heating and humidity (Extended Data Fig. 1). Dragonflies were able to live and forage successfully for weeks in this facility. The dimensions of the room were 5.5 m × 4.3 m × 4.6 m (length × width × height). Naturalistic scenery on the walls provided optic flow cues needed by dragonflies for stable flight. An array of 36 high-intensity-discharge (HID) lights was used to illuminate the room. 350 W bulbs (Philips CDM-TMW) were housed in pairs driven by a 220 Hz ballast. Flicker amplitude of the lighting system was negligible. The room was supplemented with additional UV lighting (Sylvania 350 Blacklight). The HID array produced considerable heat, both from the bulb fixtures as well as the emitted light. To allow for precise temperature control in the flight arena, the lights were segregated outside the arena in a box that enclosed the perimeter of the room at a height of 3.4 m. The inner walls of this sealed box were made of transparent acrylic. HID lights were aimed through the acrylic so they pointed upwards at roughly 45°, towards the textureless white ceiling of the flight arena. Typical illumination levels in the flight arena were approximately 10 mW cm<sup>-2</sup>. Because the light box was sealed, each of its four walls could be coupled to an air-conditioning unit that cooled the ballasts and fixtures without creating any wind within the flight area. Finally, chilled water pipes concealed within the ceiling of the flight arena provided passive heating/cooling of the interior space, allowing temperature and humidity control (31 °C, 55%), comparable to summer conditions. A single foraging platform (0.6 × 0.6 m, raised 0.3 m off the floor) was placed in the centre of the flight arena. Insect prey was attracted to the platform by placing dishes of fly food nearby. The prey in turn attracted dragonflies to the platform to forage. In a second set of experiments, we installed a custom-made artificial prey delivery system above the platform and a nylon netting covering the feeding area (1.5 × 1.5 × 2 m). The netting kept the experimental dragonflies (with retroreflective markers) in the foraging area.

**Centroid tracking system.** Two Photron SA1 cameras were mounted at a height of 3.4 m on orthogonal walls, and were aimed at the perching platform. This geometry provided a field-of-view of approximately 1 m<sup>3</sup>. Flights were filmed at 1,000 frames-per-second (fps), with a 1/2,000 s shutter speed. Identification of dragonflies and prey in the videos was made straightforward by the high contrast against the grey acrylic background provided by the platform. The software DLTdv3 (ref. 30) was used to calibrate the cameras (via the positions of nine static markers mounted on the perching platform), track the centroids of dragonfly head and prey body within each camera's view, and, from this data, reconstruct the time-varying three-dimensional (3D) position of each participant. The dragonfly head centroid was used, rather than its body, as it was less occluded by wings, etc., and hence could be tracked more reliably. Camera residuals were approximately 150 µm at the height of the perching platform, confirmed by tracking beads moved on a calibrated micromanipulator. Tracking uncertainty was less than 500 µm at all locations in the volume.

**Micro retroreflective markers.** We used motion-capture techniques and novel reflective markers to measure the full 3D head and body kinematics in foraging dragonflies. At the core of the system are custom-made retroreflective markers (750 µm and 1,000 µm diameter) that are light enough to be carried by a flying dragonfly (1.2 mg for the 750 µm marker) but reflective enough to be visualized and tracked by a camera array at a distance of up to 4 m. The retroreflectors are ball lenses (Knight Optical, Lasfn-35), with a refractive index of 2.0—optimal for a spherical retroreflector. The ball lenses were coated with an antireflective layer matched to the 790 nm spectrum of the motion-capture strobes. Moulds were fabricated to isolate one hemisphere of the ball lenses, allowing it to be mirrored with silvering solution (Angel Gilding, Oak Park, IL). The final retroreflector was a refractor-reflector pair, with the refractive outer surface focusing light onto the mirrored back surface, where it would reflect back to the outer lens and refract outwards on the original line-of-sight path of the light source. Retroreflectors were attached to the dragonfly using UV cured superglue (Loctite 4305), either as isolated 750 µm markers on each side of the head, or as a 3-marker (1,000 µm each) rigid carbon frame on the body (the body-marker frame, mass 7 mg; Fig. 2a and Extended Data Fig. 3a, b). The rigid frame fixed the 3-marker geometry as a right-angled triangle, and could be easily tracked by the cameras. It also allowed a frame-by-frame 3D reconstruction test of calibration accuracy (see below). The spherical retroreflectors had 100°–150° of functional

coverage, and great effort was made to ensure that the attached marker orientations were optimal for camera visibility during typical body and head rotations made by a flying dragonfly. The attachment of the body marker frame did not appear to affect foraging. The effect of head marker attachment varied across individuals. Some dragonflies did not forage, and displayed excessive grooming. Other dragonflies sporadically attempted foraging but were unsuccessful. These failures probably reflected either a mechanical disturbance from the markers (for example, an imbalance in their centre of mass), or a disturbance to sensory hairs. For roughly half the animals, there was no detectable effect on flight performance and the dragonflies continued to capture rapidly moving prey successfully (compare Fig. 1c to Fig. 2d). It is possible that placement of markers on the head created a slight imbalance, and hence the quality of foveation and disturbance cancellation might be even better in a dragonfly without head markers.

**Motion-capture system.** An 18-camera commercial motion-capture system running at 200 Hz (shutter: 1/3,000 s) was customized to track the retroreflective markers (Raptor-4; Motion Analysis Corp.). The original strobes on cameras were replaced with 250 790-nm LEDs. Infrared (IR) notch filters (Chroma, part et790/30\_52CR) were used to ensure that all the light the cameras received was driven by the IR strobes and not the visible HID lighting used to illuminate the flight arena. Commercial motion-capture software (Cortex, Motion Analysis Corp.) and custom software written in MATLAB (MathWorks, Inc.) were used to identify the 3D marker data based on the known distances between the body markers (6 mm, 8 mm, 10 mm) and, separately, between the two head markers (roughly 5 mm). Camera residuals were typically 150–200 µm over a 1 m × 1 m × 2 m volume. Using a calibrated micrometer, we found displacements as small as 50 µm could be reliably tracked for stationary beads. The camera calibration drifted over time, and we recalibrated the system as often as needed to maintain the average tracking residual of less than 200 µm (for the 1,000 µm markers). The rigid body-marker frame allowed us to assess the tracking accuracy during flight by analysing the jitter in the measured length of the carbon fibre rods; this was found to be 100–200 µm, consistent with the camera residuals. Because the cameras track the marker centroids in hardware, the data load of the system is small. This permits long flight durations to be recorded, as well as the possibility of closed-loop control.

**Artificial prey delivery system.** To elicit foraging flights with controlled prey statistics, we designed a system to deliver artificial prey to the perched dragonflies (Extended Data Fig. 1). A 2 mm retroreflective bead on transparent fishing line was actuated by a computer-controlled, height-adjustable pulley system. The bead was moved at either constant speed (0.2–1.5 m s<sup>-1</sup>) or with one or more pre-programmed speed variations mid-trajectory (amplitude 20–80% of the initial speed); these metric speeds, and the bead size, were based on real prey data. In all cases the artificial prey's direction of motion was constant. For analysis purposes, we classified as constant-speed prey the trials in which the maximum deviation of prey speed from its mean value was <10% (in the period following dragonfly take-off). All the other trials were classified as variable-speed prey. The 3D position of the bead was tracked with the motion-capture system described above. Dragonflies took off after approximately 10% of the artificial prey we presented, and approximately 50% of these pursuits led to successful capture. The low take-off rates may be due to the strongly linear motion of the artificial prey which rarely exists in nature. For the purpose of this study, such prey motion was needed to rule out the parallel navigation model and to analyse foveation quality for fully predictable prey motion.

**Data processing and analysis.** All data processing and analysis was performed in MATLAB (Mathworks), unless otherwise specified. A sufficient number of dragonflies and flights were collected for each condition such that all results could be reproduced robustly. No statistical method was used to predetermine sample size.

**Pre-processing and analysis of real prey trials (centroid tracking).** Three-dimensional trajectories of dragonfly (head centroid) and *Drosophila* (body centroid) were reconstructed as described above with DLTdv3 (ref. 30). Missing samples (for example, due to visual obstruction) were patched taking the corresponding samples from a cubic spline fit of the overall trajectory (gaps >45 ms were not used). Trajectories were then filtered with a finite impulse response (FIR) low-pass filter with bandwidth of approximately 20 Hz (order 160, cutoff frequency 10 Hz) and corrected for the linear phase shift. This choice of filter removes not only high-frequency noise but also the small effects of individual wing beats on the trajectories, which are not addressed in this study (prey motion relative to the dragonfly does not occur in this bandwidth and hence could be filtered out by the dragonfly); the absence of phase distortion guarantees that the manoeuvring timing, which is a main focus of this research, is not affected. To minimize edge effects at the moment of capture (since the prey disappears here and the time series ends), the raw data were extrapolated (cubic spline) by the half-width of the FIR filter. Numerical derivatives were computed using a Savitsky–Golay filter of length 11 and order 4, and corrected for linear phase shift.



We defined the dragonfly take-off time as the first sample at which the vertical speed of the dragonfly increased above  $0.15 \text{ m s}^{-1}$ . Capture time was defined as the last sample at which the prey position was available, for successful trials, or the sample of minimal distance, for unsuccessful trials. 140 real prey flights (110 successful captures, 30 close misses with minimal distance  $< 40 \text{ mm}$ ) from 18 dragonflies were used in the analysis; the median flight duration in these data was 310 ms (maximum  $\sim 1,000 \text{ ms}$ ). In the analysis of range vector correlation (Fig. 1f, g), we excluded times preceding dragonfly take-off, when changes in the range vector only depend on the prey trajectory.

To analyse the timing relation between prey and dragonfly steering in successful trials, we identified as significant events the peaks in acceleration magnitude that were larger than the typical acceleration, given by the median across all trials ( $5.46 \text{ m s}^{-2}$  for prey,  $10.86 \text{ m s}^{-2}$  for dragonfly). We excluded dragonfly acceleration peaks within the first 25 ms after take-off, that arise systematically due to the lift-off manoeuvre, as well as those occurring within the last 20 ms before capture, as they reflected the prey grasping, non-interception portion of the flight (Supplementary Video 1). We identified 155 prey steering events and 190 dragonfly steering events across the 110 successful trials.

To identify how many of the dragonfly steering events were likely to be direct responses to prey steering events, we analysed the lag distribution between pairs of prey-steering and dragonfly-steering events occurring within the same trial ( $n = 332$  event pairs). We also computed the control distribution expected if the timing of prey and dragonfly steering events were independent of each other. The control distribution was estimated by replacing the prey steering events within each trial with an equal number of events uniformly distributed in time, and averaging over 1,000 iterations. The control distribution was then subtracted from the empirical lag distribution. The results of the steering timing analysis (Fig. 1e) were robust to changes in the criteria used in selecting the steering events (that is, minimum threshold for acceleration peaks).

**Pre-processing of retroreflective marker trajectories (motion capture).** The identified 3D marker trajectories were obtained as described above using Cortex (Motion Analysis Corp.) and MATLAB (Mathworks). Missing samples (from untracked markers, for example, due to occlusion) were patched via a cubic spline fit of the overall trajectory, excluding gaps  $> 25 \text{ ms}$ . Trajectories were then filtered with an FIR low-pass filter with bandwidth of about 20 Hz (order 32, cutoff frequency 10 Hz, comparable to the filter used earlier for centroid tracking) and corrected for the linear phase shift. To minimize edge effects at the moment of capture (since the prey disappears here and the time series ends), the raw data were extrapolated (cubic spline) by the half-width of the FIR filter. Numerical derivatives were computed using a central difference filter of order 9 and also corrected for the linear phase shift.

**Identification of head joint location.** The connection between the dragonfly's head and its thorax occurs over a very small, point-like area<sup>31</sup>, which we refer to as the head joint. To reduce the number of markers carried by the dragonfly's head, we numerically estimated the location of the neck joint from the head and body markers. We estimated its position by adapting a technique developed to determine the centre of rotation of a ball joint<sup>32</sup>. The method numerically fit the joint as the centre of a collection of concentric spheres, each of which describes a sample rotation. To evaluate its performance, we applied the method to a rotating pin, and recovered the point of rotation with approximately  $150 \mu\text{m}$  accuracy.

To apply the method to dragonflies, we used data from the saccadic head movements made by each dragonfly before engaging in prey pursuit (Supplementary Video 3). These head saccades are rigid rotations of the head about the joint while the body remains stationary. For each dragonfly with head markers, at least 48 head saccade trajectories were used to infer the head joint location relative to the body markers. In each interception trial, the frame-by-frame position of the head joint was then reconstructed from that of the body markers via this transformation. One dragonfly was excluded from the analysis of foveation as it had insufficient head saccade data to accurately compute the neck joint.

**Head and body kinematics.** To allow cross-animal comparison, we mapped the body and head reference frames to stereotypical anatomical features on each dragonfly (Extended Data Fig. 3). This mapping was performed immediately after the retroreflective markers were attached, using a rig comprised of two colour HD cameras, or a single camera paired with a mirror. From the two close-range images of the dragonfly, we reconstructed the 3D relationships between markers and dragonfly anatomical features using methods described previously<sup>30</sup>. The marker head frame was rotated into a fixed foveal head frame such that the  $x$  axis (roll) pointed towards the foveal centre (mean elevation of prey on head across all trials) used by

each dragonfly. The body marker frame was pitched downwards into a body frame such that the  $x$  axis (roll) corresponded to the longitudinal axis of the dragonfly's body (Fig. 2c). See Extended Data Fig. 3c for details. Euler angles representing head or body orientations (for example, Fig. 3), were defined based on a  $z$ - $y$ - $x$  Tait–Bryan convention; the first two angles correspond to the azimuth and elevation of the longitudinal body axis ( $\mathbf{x}_B$ ) with respect to the world reference frame. Angular velocities (Fig. 5 and Extended Data Fig. 5) were computed from numerical differentiation of the quaternion representations of such orientations.

**Analysis of artificial prey trials.** Data for the body dynamics steering analysis (Figs 1g and 3) were time-aligned to capture and required only body markers. This analysis included all the trials in which the dragonfly and prey trajectories could be tracked at least until their minimum distance was reduced to below 40 mm. This threshold was typically reached approximately 30 ms before capture; the marker samples after this time window were sometimes lost due to occlusions when the body rotated to grasp the prey. This data set comprised 40 trials (26 constant-speed prey, 14 variable-speed prey) from 9 dragonflies. The dragonfly trajectories were computed at the head joint, which represents the origin of the head reference frame relevant to foveation, and allows comparison with the head centroid trajectories computed in the Photron data set (Fig. 1). Head joint trajectories were computed from those of the body markers as described above, using animal-specific transformations when available (5 head-marked dragonflies) or average transformations otherwise (4 dragonflies without head markers). Take-off and capture times were defined as in real prey trials.

The prey foveation and head-control analysis (Figs 4 and 5) required both head and body marker data. The foveation data set comprised 37 flights (21 constant-speed prey, 16 variable-speed prey, from 5 dragonflies). The data were time-aligned to take-off and used 50 ms of data before take-off as foveation was already established and compensating for preparatory body movements. The 50 ms immediately before capture were omitted from analysis as the large body rotation initiated in this time interval (to grasp the prey) shifts the prey to outside the field-of-view of the dorsal eye.

In Fig. 4d we showed the actual prey trajectories as seen in the anatomical head frame, assuming the prey is a point. This is justified by the fact that at an average distance of 400 mm, a 2 mm prey subtends  $0.3^\circ$ —slightly smaller than the approximate angular size of a single ommatidium in the high-resolution portion of the eye<sup>19</sup>. In Fig. 4a–c we showed the trajectories that would have been produced by each drift source alone: rotation of the body, translation of the prey relative to dragonfly, and rotation of the head relative to the body. Each trajectory was computed by numerically integrating (Simpson's rule) the individual terms in the drift rate equations (Extended Data Fig. 4), given the prey location and our measurements of head rotation, body rotation, etc. Numerical integration of the sum of the three drift terms produced trajectories indistinguishable from those obtained by direct measurements (Fig. 4d). The root mean square (RMS) estimates of prey jitter were computed by combining all the trials, after removing the trial-specific mean location of the prey during interception.

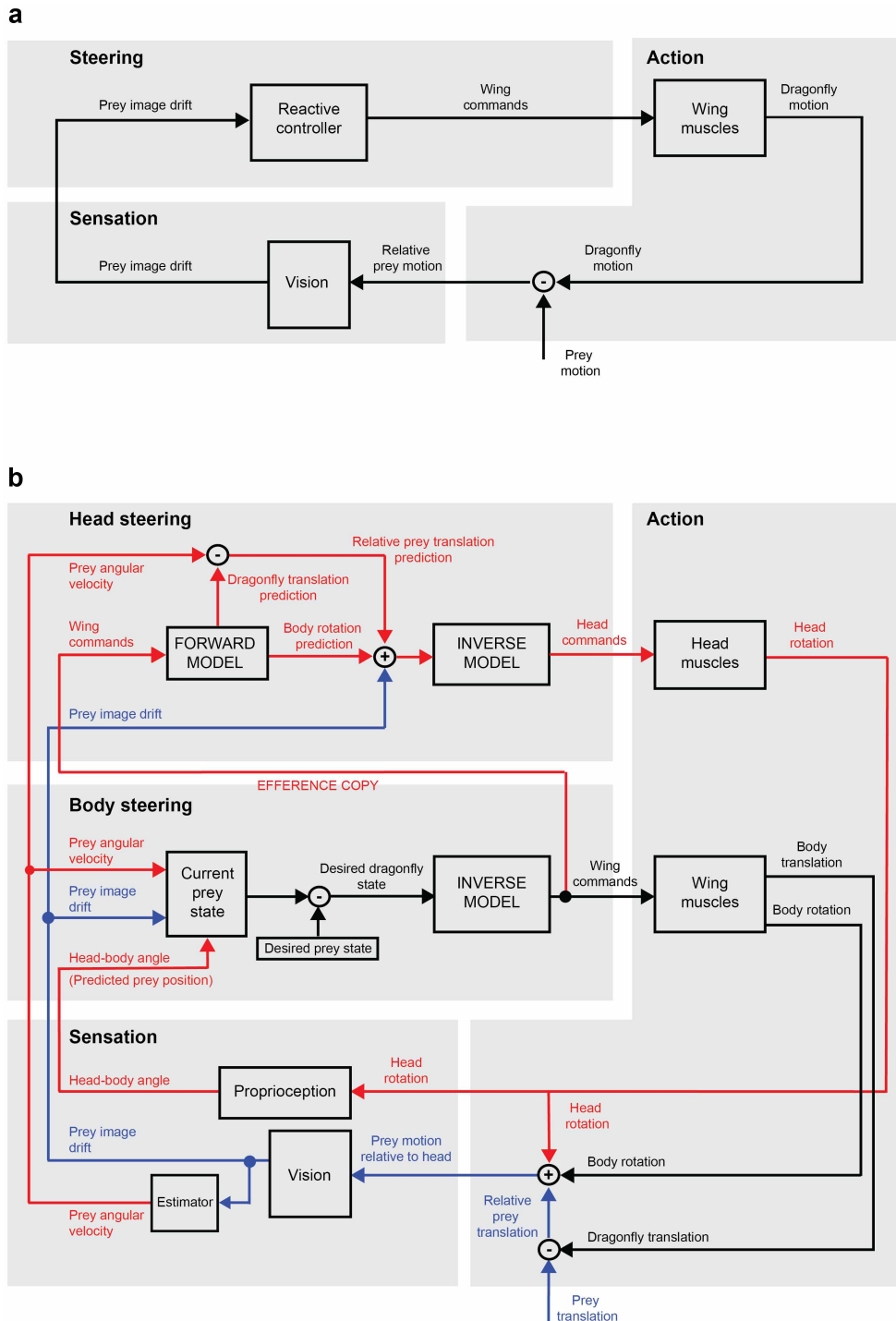
In Fig. 5, the quantities plotted are those in the equations of Extended Data Fig. 4, except that the azimuthal terms were multiplied by the cosine of the elevation (a natural rescaling that makes azimuthal and elevational rates comparable to each other). To compute the residual head rotations that compensate for translational prey-image drift, we added the body rotation to the head rotation. This removes any body rotation prediction from the head rotation signal and leaves a combination of prediction error, prey translation compensation, and other sources of noise (see Fig. 5g–k). The multi-dimensional cross-correlation functions showed in Fig. 5 were computed using a vector comprised of azimuth and elevation (but not the range direction), on a trial-by-trial basis. Margin of errors ( $\pm 1$  standard deviations) on the slopes of the linear fits were computed from bootstrapping (10,000 iterations). The percentages of foveation disturbance reduction (Fig. 5) were computed from the median, over all trials, of the ratio between prey-image drift rate magnitude before and after cancellation (body rotation + relative translation versus body rotation + relative translation + head rotation).

30. Hedrick, T. L. Software techniques for two- and three-dimensional kinematic measurements of biological and biomimetic systems. *Bioinspir. Biomim.* **3**, 034001 (2008).
31. Gorb, S. N. Evolution of the dragonfly head-arresting system. *Proc. R. Soc. B* **266**, 525–535 (1999).
32. Chang, L. Y. & Pollard, N. S. Constrained least-squares optimization for robust estimation of center of rotation. *J. Biomech.* **40**, 1392–1400 (2007).



**Extended Data Figure 1 | Indoor flight arena enables recording of dragonfly foraging behaviour.** Flight arena is a  $5.5 \text{ m} \times 4.3 \text{ m} \times 4.6 \text{ m}$  room with controlled lighting ( $10 \text{ mW cm}^{-2}$ , 220 Hz illumination), heating ( $31^\circ \text{C}$ ) and humidity (55%), comparable to outdoor summer conditions. Naturalistic images on the walls provide dragonflies with the optic flow needed for flight stability. **a–c**, A raised platform ( $0.6 \text{ m} \times 0.6 \text{ m}$ ) (a), used by the dragonflies as a foraging perch, is placed at the centre of the room, within the filming volume of two high-speed cameras (b, Photron SA1, 1,000 fps,  $150 \mu\text{m}$  camera

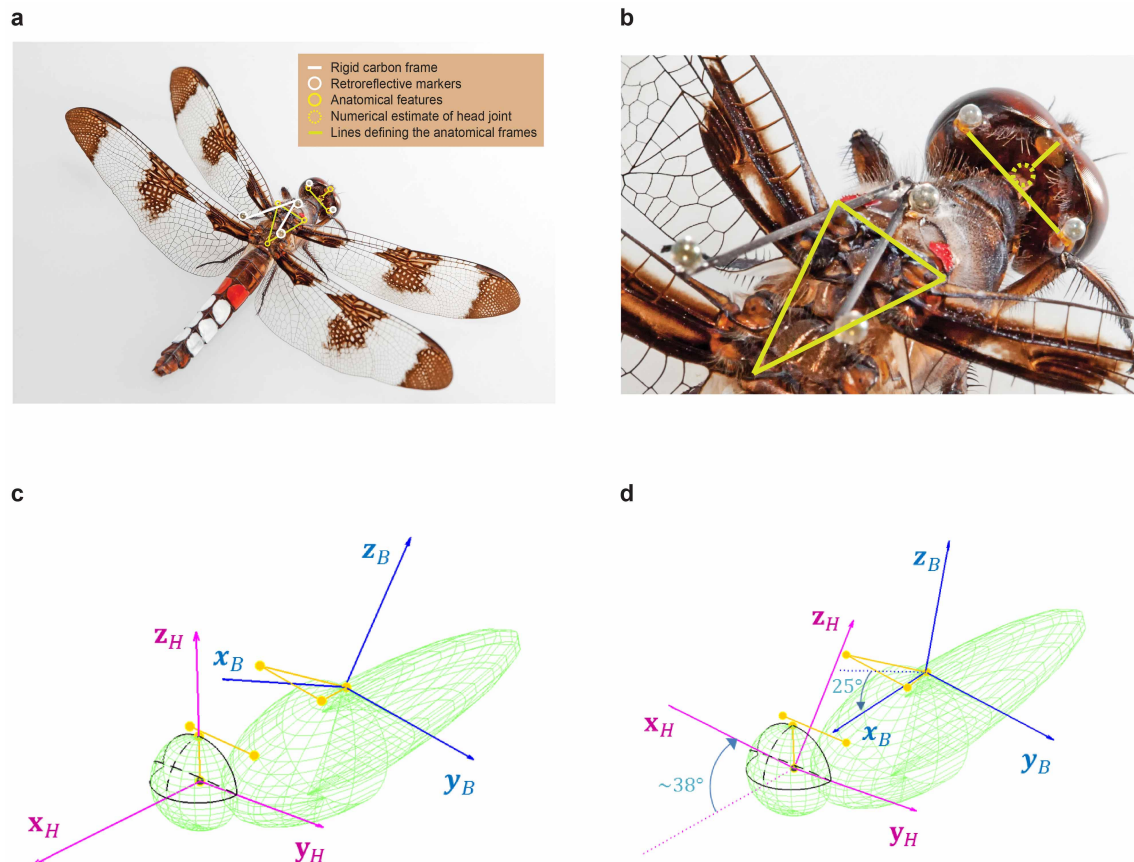
residuals) and 18 infrared motion-capture cameras (c, Motion Analysis Corp. Raptor-4, 200 fps,  $150 \mu\text{m}$  camera residuals). Insect prey (*Drosophila virilis*) are stocked in the arena and attracted to the platform by fly food dishes (not shown). **d**, An artificial prey delivery system is installed above the platform; this comprises a computer-controlled, height-adjustable pulley system (dashed yellow line) which moves a 2-mm retroreflective bead on transparent fishing line. A netted enclosure ( $1.5 \text{ m} \times 1.5 \text{ m} \times 2 \text{ m}$ ) keeps the instrumented dragonflies in the filming area during experiments.



**Extended Data Figure 2 | Classical reactive steering versus model-driven interception steering in the dragonfly.** **a**, Classical reactive steering: prey-image drift directly drives steering commands to maintain a desired strategy, such as parallel navigation (Fig. 1a). Dragonfly motion and prey motion alter prey angular velocity and position, thereby driving the next round of steering. **b**, Conceptual model summarizing the proposed control architecture used by the dragonfly—a predictive steering pathway (red) and a visually driven reactive steering pathway (blue) to explain our results. Steering of the body is driven by prey position, prey angular velocity and the orientation of the body relative to the zenith (Fig. 3). We conjecture that these variables (the current prey state) are compared to a desired prey state—prey overhead, with the dragonfly body axis aligned to the prey's flight path—and the state error is used to produce a wing command that yields movement of the dragonfly's body. An inverse model may be used to transform the desired dragonfly state (in sensory coordinates) into a motor command; future studies will be needed

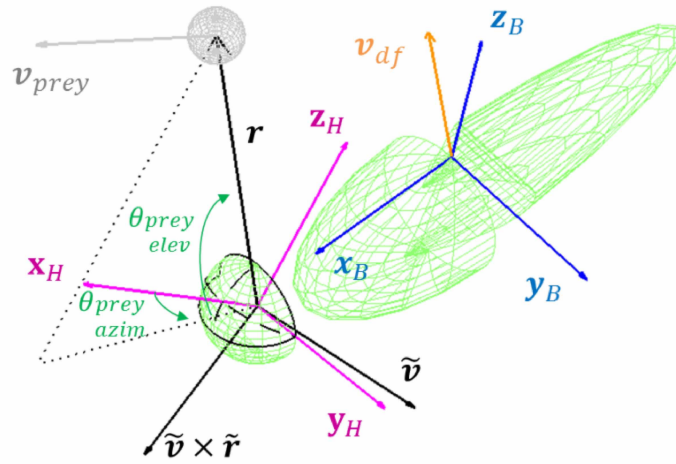
to confirm this. Head steering to stabilize the prey in the fovea (Fig. 4d) is driven by prey angular velocity and an efference copy of the wing command signal. These variables yield a forward-model prediction of expected foveal drift, which is passed through an inverse model to generate compensatory head rotations (Figs 4c and 5). Accurate foveation can be achieved in the absence of unexpected prey motion (for example, blue trial, Fig. 4d), eliminating prey-image drift on the head and hence activity of the visual reactive pathway. In this regime, only the predictive pathway drives dragonfly steering. The predicted prey position can be inferred from the head-body angles (encoded proprioceptively, or estimated via efference copy of neck commands), whereas prey angular velocity must be available as an internal state variable. When prey-image drift on the retina detects unexpected prey motion, the reactive pathway introduces corrective steering of both body and head (Fig. 1e) and updates the estimate of prey angular velocity.





**Extended Data Figure 3 | Definition of anatomically based head and body reference frames.** **a**, Position of the retroreflective markers is mapped to stereotypical anatomical features, which are used along with the numerically estimated head joint to define anatomical reference frames. **b**, Anatomical features include three points on the thorax (leftmost and rightmost tips of dorsal carina of episternum, painted in red for ease of identification, and centre of hind wing scotellum) and three points on the back of the head (upper tip of midline and two laterally symmetrical points, for example, the tips of yellow coloured bands underneath the head markers). **c**, A preliminary body frame is defined based on the triangle formed by the three anatomical points on the thorax, with its origin at the rear vertex of the triangle. From this vertex, the roll axis ( $x_B$ ) is defined as the direction to the midpoint between the other two vertices, the yaw axis ( $z_B$ ) as the direction orthogonal to the triangle, and the pitch axis ( $y_B$ ) as the direction orthogonal to the other two. A preliminary head

frame, with origin at the head joint, is defined based on the head anatomical features. From the joint, the yaw axis ( $z_H$ ) is the direction to the top anatomical point. The pitch axis ( $y_H$ ) is the direction pointing from the right to the left anatomical points on the head, orthogonalized relative to the yaw axis. Finally, the roll axis ( $x_H$ ) is orthogonal to the other two. In the schematic, the black solid contour on the head denotes its dorsal part, defined by this head frame. **d**, The final body frame is obtained applying a 25° pitch correction to the preliminary one, to compensate for the natural upward slope of the anatomical triangle. The corrected body axis roughly corresponds to the thorax–abdomen line (Fig. 2c). The final ('foveal') head frame is instead obtained by applying a pitch correction to the preliminary one so that the roll axis matches the average prey-image elevation during pursuit. This provides an estimate for the centre of the high-acuity region of the eye, and varies slightly across animals (38° average).



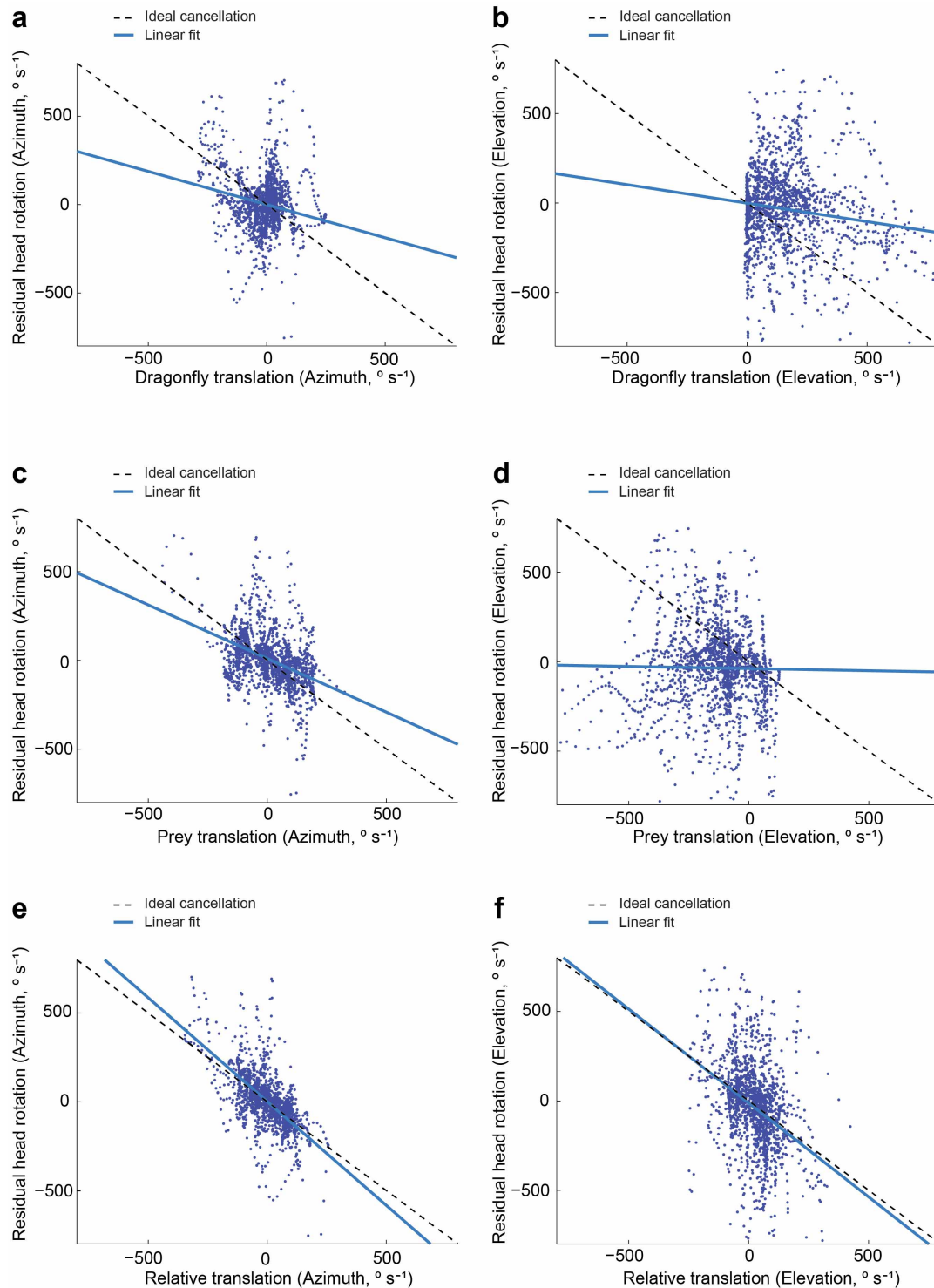
$$\dot{\theta}_{prey,azim} = \left[ \omega_{Hrel}^{\tilde{v} \times \tilde{r}} + \omega_{Babs}^{\tilde{v} \times \tilde{r}} + \frac{-v_{df}^{\tilde{v}} + v_{prey}^{\tilde{v}}}{|r|} \right] \frac{1}{\cos(\theta_{prey,elev})}$$

$$\dot{\theta}_{prey,elev} = \omega_{Hrel}^{\tilde{v}} + \omega_{Babs}^{\tilde{v}} + \frac{v_{df}^{\tilde{v} \times \tilde{r}} - v_{prey}^{\tilde{v} \times \tilde{r}}}{|r|}, \tilde{r} \triangleq \frac{\mathbf{r}}{|r|}, \tilde{v} \triangleq \frac{\mathbf{z}_H \times \mathbf{r}}{|\mathbf{z}_H \times \mathbf{r}|}$$

$$\text{drift rate} = \text{head rotation} + \text{body rotation} + \text{relative translation (dragonfly \& prey)}$$

**Extended Data Figure 4 | Analytical decomposition of prey-image drift on the head.** The angular position of the prey image on the head is obtained by projecting the range vector ( $\mathbf{r}$ ) on the head reference frame ( $\mathbf{x}_H, \mathbf{y}_H, \mathbf{z}_H$ ); we define the azimuthal position as  $\theta_{prey,azim}$  and the elevational position as  $\theta_{prey,elev}$ . Prey-image drift ('drift', with rates  $\dot{\theta}_{prey,azim}, \dot{\theta}_{prey,elev}$ ) can be induced by changes in the orientation of the head, and by changes in the direction of the range vector (caused by translation of the prey relative to the dragonfly). Changes in head orientation can be quantified as the sum of the absolute angular velocity of the body ( $\omega_{B,abs}$ , 'body rotation') and the relative angular velocity of the head with respect to the body ( $\omega_{H,rel}$ , 'head rotation'). The contribution of relative translation to the drift rate equals the difference between dragonfly ( $\mathbf{v}_{df}$ ) and prey ( $\mathbf{v}_{prey}$ ) velocities, scaled by the distance to the

prey ( $|r|$ ). The relevant dragonfly velocity for foveation is that at the origin of the head frame; for graphical clarity, we have shown the velocity at the origin of the body reference frame ( $\mathbf{x}_B, \mathbf{y}_B, \mathbf{z}_B$ ) instead. The notation  $\mathbf{a}^b$  stands for the projection of vector  $\mathbf{a}$  in the direction of vector  $\mathbf{b}$ . Prey movement in the head frame ( $\mathbf{x}_H, \mathbf{y}_H, \mathbf{z}_H$ ) (Fig. 4) will generally produce motion simultaneously along all three axes. We analysed foveation within the coordinate system defined by ( $\tilde{\mathbf{r}}, \tilde{\mathbf{v}}, \tilde{\mathbf{v}} \times \tilde{\mathbf{r}}$ ) (Fig. 5) because it decouples the directions functionally relevant for foveation. Rotations about  $\tilde{\mathbf{v}}$  produce purely elevational drift, whereas rotations about  $\tilde{\mathbf{v}} \times \tilde{\mathbf{r}}$  produce purely azimuthal drift. Translations in the range vector direction and rotations about that axis only change the angular size and orientation of the prey on the eye but cause no drift in position.



**Extended Data Figure 5 | Dragonfly head movements compensate for both dragonfly translation and prey translation.** **a–f**, Scatter plots of residual head rotation against individual prey and dragonfly translational components, and against their sum (21 constant-speed artificial prey flights). Cancellation of dragonfly translation alone (**a**, azimuth: slope  $m = -0.38 \pm 0.07$ ,  $r^2 = 0.03$ ;

**b**, elevation:  $m = -0.21 \pm 0.04$ ,  $r^2 = 0.02$ ) and prey translation alone (**c**, azimuth:  $m = -0.60 \pm 0.04$ ,  $r^2 = 0.14$ ; **d**, elevation:  $m = -0.02 \pm 0.04$ ,  $r^2 < 0.01$ ) is largely inferior to the cancellation of the relative prey translation (**e**, azimuth:  $m = -1.17 \pm 0.05$ ,  $r^2 = 0.38$ ; **f**, elevation:  $m = -1.05 \pm 0.08$ ,  $r^2 = 0.14$ ).



# Impact jetting as the origin of chondrules

Brandon C. Johnson<sup>1</sup>, David A. Minton<sup>2</sup>, H. J. Melosh<sup>2</sup> & Maria T. Zuber<sup>1</sup>

Chondrules are the millimetre-scale, previously molten, spherules found in most meteorites<sup>1</sup>. Before chondrules formed, large differentiating planetesimals had already accreted<sup>2</sup>. Volatile-rich olivine reveals that chondrules formed in extremely solid-rich environments, more like impact plumes than the solar nebula<sup>3–5</sup>. The unique chondrules in CB chondrites probably formed in a vapour-melt plume produced by a hypervelocity impact<sup>6</sup> with an impact velocity greater than 10 kilometres per second. An acceptable formation model for the overwhelming majority of chondrules, however, has not been established. Here we report that impacts can produce enough chondrules during the first five million years of planetary accretion to explain their observed abundance. Building on a previous study of impact jetting<sup>7</sup>, we simulate protoplanetary impacts, finding that material is melted and ejected at high speed when the impact velocity exceeds 2.5 kilometres per second. Using a Monte Carlo accretion code, we estimate the location, timing, sizes, and velocities of chondrule-forming impacts. Ejecta size estimates<sup>8</sup> indicate that jetted melt will form millimetre-scale droplets. Our radiative transfer models show that these droplets experience the expected cooling rates of ten to a thousand kelvin per hour<sup>9,10</sup>. An impact origin for chondrules implies that meteorites are a byproduct of planet formation rather than leftover building material.

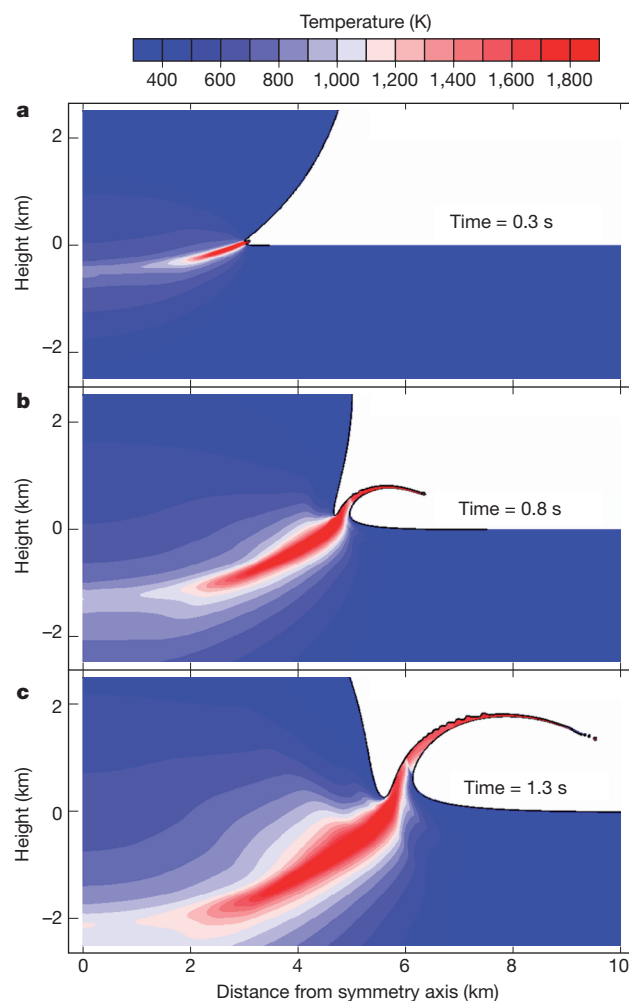
Recent work shows that the iSALE hydrocode<sup>11,12</sup> is capable of modelling the extreme process of impact jetting<sup>7</sup>. While previous models focused on impacts between strengthless fluid bodies<sup>7</sup>, here we determine the effect that porosity and material strength have on the jetting process (Methods). Using iSALE we simulate vertical impacts of initially fractured dunite impactors of 1%, 10%, and 25% porosity and 10-km diameter on flat targets at impact velocity  $v_{\text{imp}} = 1\text{--}6\text{ km s}^{-1}$  stepping up by  $0.5\text{ km s}^{-1}$  (Methods). Later, we scale our results to larger impactor sizes using hydrodynamic similarity<sup>13</sup>. We find that no melted material (Methods) ejected above escape velocity is resolved for impacts with  $v_{\text{imp}} < 2.5\text{ km s}^{-1}$ , assuming that  $v_{\text{imp}}/v_{\text{esc}}$  is between 0.5 and 2, where  $v_{\text{esc}}$  is the escape velocity. During the impact simulation shown in Fig. 1, a total mass of material equivalent to  $\sim 1\%$  of the impactor's mass is melted and ejected at higher than escape velocity for an assumed  $v_{\text{imp}}/v_{\text{esc}} = 1$ .

As shown in Fig. 1, jetting melts and ejects near-surface material at high velocity. Thermal modelling and palaeomagnetism of chondrites indicate that some differentiated planetesimals had outer shells of undifferentiated material<sup>14</sup>. If this near-surface material is undifferentiated, jetting will produce chondrules with primitive compositions<sup>15</sup>. This is in contrast to other models that rely on splashing during low-velocity ( $v_{\text{imp}} \approx 10\text{--}100\text{ m s}^{-1}$ ) collisions between already molten planetesimals<sup>16,17</sup>. It is doubtful that ejection of previously molten and differentiated<sup>2</sup> material would produce chemically unfractionated chondrules<sup>18</sup>. Moreover, collisional splashing creates 'droplets' that are approximately 40 m in diameter (Methods). Previous size estimates<sup>16</sup>, which agreed with observed chondrule sizes, neglected the effect of decompression heating<sup>19</sup>.

Using the GAME Monte Carlo accretion code<sup>20</sup> we are able to determine where and when chondrule-forming impacts will occur. We model a typical minimum mass solar nebula (MMSN) and a three times more massive nebula (3MMSN), both of which extend from 0.4 astronomical units (AU) to 4 AU (ref. 20). Initially, solid bodies have a main-belt-like

size frequency distribution between 100 km and 1,000 km in diameter<sup>21</sup> (Methods). In addition to these bodies, our models include the eccentricity damping effect of nebular gas<sup>20</sup>. As Fig. 2 shows, chondrule-forming impacts occur about  $10^3\text{--}10^4\text{ yr}$  into the simulation. Assuming that the initial conditions of our model correspond to the time of CAI formation  $t_{\text{CAI}}$ , this timing of chondrule formation is consistent with the age of the oldest chondrules,  $t_{\text{CAI}} \pm 0.4$  million years (Myr) (ref. 22).

The target bodies for impacts with  $v_{\text{imp}} < 2.5\text{ km s}^{-1}$  are planetary embryos more massive than the Moon. This explains how impacts,



**Figure 1 | Jetting of melted material during an accretionary impact.** A time series showing a 1% porous projectile 10 km in diameter striking a target at  $3\text{ km s}^{-1}$ . Material is coloured according to its temperature at the time shown. The origin is the collision site. In **a** the jet is just beginning to form, 0.3 s into the impact. Panel **b** shows a well-formed hot jet 0.8 s into the impact. Panel **c** roughly shows the end of jetting, when the projectile penetrates halfway into the target. The fastest ejecta have a velocity of about  $6\text{ km s}^{-1}$ , or twice the impact velocity. The figure was produced using iSALEPlot.

<sup>1</sup>Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. <sup>2</sup>Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, 550 Stadium Mall Drive, West Lafayette, Indiana 47907, USA.

which eject much more solid material than melted material, can produce chondrule-rich chondrites. The only material that escapes the target bodies is a high-speed mixture of melted surface material (nascent chondrules) and lightly shocked cold proto-matrix (matrix is the material in which chondrules are embedded, and shocking refers to the increase in material temperature and pressure by passage of a shock wave). Most of the lower-speed solid ejecta are retained because of the target bodies' large escape velocity. The ejected chondrules and dust are decelerated to low relative velocities by lingering nebular gas, and preferentially accrete onto smaller planetesimals, which collectively have a larger surface area than more massive bodies<sup>23</sup>.

Chondrule formation stopped sometime around 5 Myr after  $t_{\text{CAI}}$  (refs 1 and 22). In our models, large impacts continue to occur even after 5 Myr. However, chondrules are ejected at velocities above about  $2.5 \text{ km s}^{-1}$  and will have dynamically excited orbits. Without significant gas present to reduce the eccentricities and inclinations of their orbits, chondrules will break up as they collide with one another or larger bodies at velocities comparable to their ejection velocity of a few kilometres per second. Astronomical observations suggest that the mean lifetime of a primordial protoplanetary disk is about 3 Myr (ref. 24). Melt ejected below escape velocity is bound to the target bodies, which are too large to be disrupted by impacts, and must be dynamically removed from the asteroid belt<sup>25</sup>.

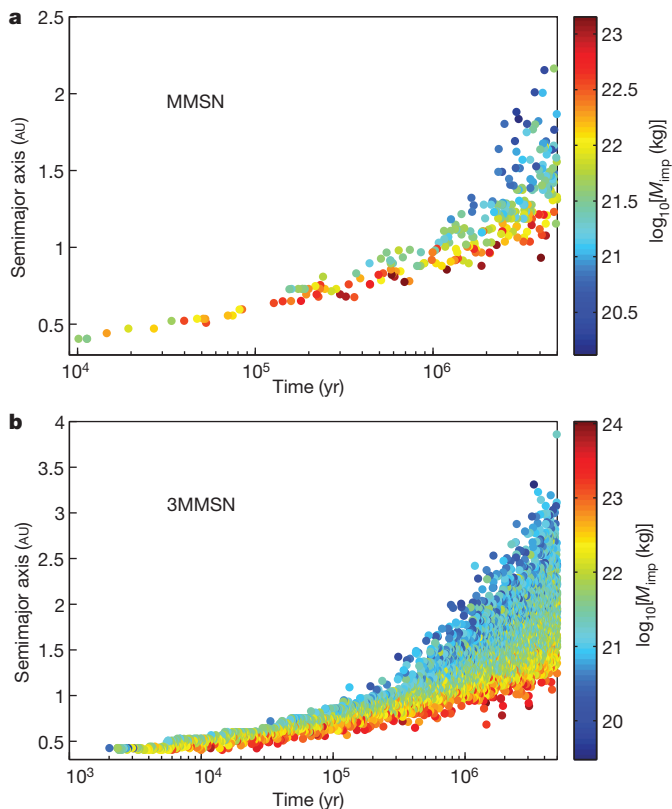
The earliest chondrule-forming impacts occur closer to the Sun and the position where chondrule formation occurs moves outward with time (Fig. 2). 5 Myr into accretion, the outermost chondrule-forming impacts occur within 2.2 AU and 3.9 AU, respectively, for the MMSN and 3MMSN models (the main asteroid belt extends from  $\sim 2$  AU to 3 AU). With masses of  $10^{20}$ – $10^{23}$  kg, the impacting bodies range in size from nearly that of Vesta to larger than the Moon (Fig. 2a). The maximum impact velocity and  $v_{\text{imp}}/v_{\text{esc}}$  increase as time goes on (Extended Data Fig. 1). As  $v_{\text{imp}}/v_{\text{esc}}$  increases, fractionally more unmelted material is

ejected at higher than escape velocity. Thus, our model predicts that chondrites formed further away from the Sun should form later and on average be richer in matrix than those that formed closer to the Sun.

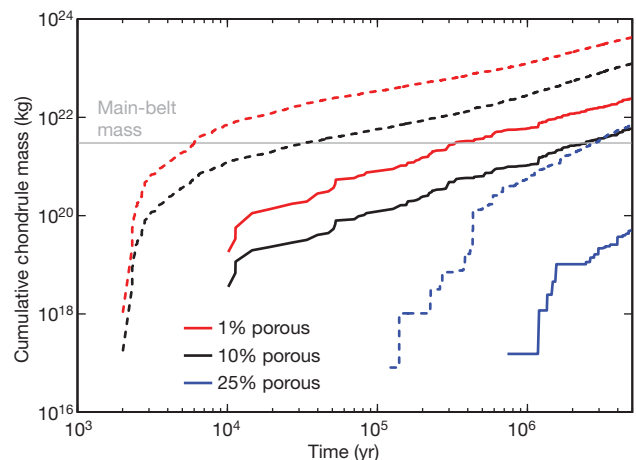
Using hydrodynamic scaling and estimates of the melt mass ejected above escape velocity as a function of  $v_{\text{imp}}/v_{\text{esc}}$  for each of our iSALE models (Methods), we estimate the mass of chondrules created by each accretionary impact occurring in our GAME simulation. The 287 chondrule-forming impacts in the MMSN model produce more than  $2 \times 10^{22}$  kg of chondrules while the 4776 chondrule-forming impacts in the 3 MMSN model make over  $4 \times 10^{23}$  kg of chondrules for the 1% porosity case (Fig. 3). Jetted mass is expected to increase significantly for impacts on more realistic curved targets<sup>7,26</sup> and may also increase when oblique impacts are considered<sup>26,27</sup>. The threshold velocity of  $2.5 \text{ km s}^{-1}$  required to produce chondrules may be an overestimate for the same reasons. A lower threshold velocity would produce more chondrules further out in the disk, as does using a different size distribution for the initial conditions of GAME (Methods). Thus, the estimates shown in Fig. 3 probably represent a minimum approximation for the total mass of chondrules that impacts can produce.

The present asteroid belt has a mass  $M_{\text{mb}} = 3 \times 10^{21}$  kg and is depleted in mass by a factor of  $\sim 1,000$  from a MMSN<sup>28</sup>. This corresponds to a numerical depletion factor of 10–100 because most of the mass was contained in large planetary embryos<sup>28</sup>. Chondrules will preferentially accrete onto smaller bodies, which have a larger collective surface area than more massive bodies. Consequently, assuming the main-belt is about one-third chondrules by mass, a successful chondrule formation mechanism must produce  $\sim 3M_{\text{mb}}$  to  $30M_{\text{mb}}$  of chondrules<sup>23</sup>. Although dynamical models indicate that  $\sim 10\%$ – $30\%$  of the asteroid belt mass may be material originally from 1.5–2 AU (ref. 25), our MMSN model only makes  $8M_{\text{mb}}$  in total. However, a 3MMSN model makes  $\sim 142M_{\text{mb}}$  of chondrules with  $11M_{\text{mb}}$  being made in the main-belt region (2–3 AU). Therefore, even with reasonable assumptions our lower limit estimates suggest that impacts produce enough chondrules to explain the current chondrule abundance.

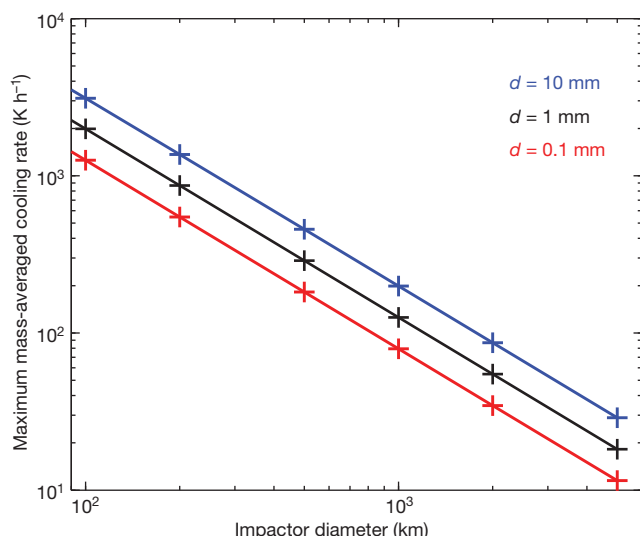
The igneous textures that chondrules exhibit imply that they cooled at rates of  $10$ – $1,000 \text{ K h}^{-1}$  (refs 9 and 10). Using a one-dimensional radiative transfer code, and a geometry that approximates that of jetted material (Methods, Extended Data Fig. 3), we determine the average cooling rates of impact-produced chondrules for a range of impactor sizes (Fig. 4, Methods). Our simulations indicate that jetted material cools at rates of  $10$ – $1,000 \text{ K h}^{-1}$  for impactors that are hundreds to thousands of kilometres in diameter. These estimates demonstrate that melted droplets jetted during large-scale accretionary impacts will exhibit the observed igneous textures of chondrules.



**Figure 2 | Timing and location of chondrule-forming impact.** Chondrule-forming impacts with velocities above  $2.5 \text{ km s}^{-1}$  for the MMSN model (a) and the 3MMSN model (b). The points are coloured according to the logarithm of the mass of the impacting body marked on the corresponding colour bar.



**Figure 3 | The cumulative mass of chondrules created by accretionary impacts.** The red, black and blue curves show model results for 1%, 10%, and 25% initial porosities. The solid curves correspond to the MMSN model and the dashed curves correspond to the 3MMSN model. The grey line acts as a guide to the eye, showing the mass of the main asteroid belt, at  $3 \times 10^{21}$  kg.



**Figure 4 | Chondrule cooling rates as a function of impactor size.** The coloured lines represent different assumed droplet diameters  $d$  within the jet, as indicated. The crosses represent actual model runs.

The chondrules found within a given chondrite type exhibit significant variations in composition<sup>1,29</sup>. Turbulence within the jet and chondrule–chondrule collisions may mix material on scales comparable to the thickness of the jet. However, the jet is composed of a mixture of target and projectile material and the ratio of target-to-projectile material changes with time<sup>7</sup>. If near surface material from the target and projectile have significantly different compositions, the chondrules produced by the impact will be chemically diverse. Expected spatial variations in chondrule number density could also contribute to the observed compositional diversity<sup>30</sup>.

An important aspect of our model is that, although chondrules accumulate to form chondrites on small bodies, the chondrules themselves are formed by impacts on much larger bodies. Because of this, only a small fraction of the mass of the terrestrial planets is processed into chondrules. Thus, we argue that chondrules are not the direct building blocks of the planets, but merely a byproduct of their accretion.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 16 July; accepted 20 November 2014.**

1. Scott, E. R. D. Chondrites and the protoplanetary disk. *Annu. Rev. Earth Planet. Sci.* **35**, 577–620 (2007).
2. Kruijer, T. S. *et al.* Protracted core formation and rapid accretion of protoplanets. *Science* **344**, 1150–1154 (2014).
3. Fedkin, A. V. & Grossman, L. Vapor saturation of sodium: key to unlocking the origin of chondrules. *Geochim. Cosmochim. Acta* **112**, 226–250 (2013).
4. Alexander, C. M. O. & Ebel, D. S. Questions, questions: Can the contradictions between the petrologic, isotopic, thermodynamic, and astrophysical constraints on chondrule formation be resolved? *Meteorit. Planet. Sci.* **47**, 1157–1175 (2012).
5. Alexander, C. M. O., Grossman, J. N., Ebel, D. S. & Ciesla, F. J. The formation conditions of chondrules and chondrites. *Science* **320**, 1617–1619 (2008).
6. Krot, A. N., Amelin, Y., Cassen, P. & Meibom, A. Young chondrules in CB chondrites from a giant impact in the early Solar System. *Nature* **436**, 989–992 (2005).

7. Johnson, B. C., Bowling, T. J. & Melosh, H. J. Jetting during vertical impacts of spherical projectiles. *Icarus* **238**, 13–22 (2014).
8. Johnson, B. C. & Melosh, H. J. Formation of melt droplets, melt fragments, and accretionary impact lapilli during a hypervelocity impact. *Icarus* **228**, 347–363 (2014).
9. Lofgren, G. & Russell, W. J. Dynamic crystallization of chondrule melts of porphyritic and radial pyroxene composition. *Geochim. Cosmochim. Acta* **50**, 1715–1726 (1986).
10. Desch, S. J., Morris, M. A., Connolly, H. C. & Boss, A. P. The importance of experiments: constraints on chondrule formation models. *Meteorit. Planet. Sci.* **47**, 1139–1156 (2012).
11. Collins, G. S., Melosh, H. J. & Ivanov, B. A. Modeling damage and deformation in impact simulations. *Meteorit. Planet. Sci.* **39**, 217–231 (2004).
12. Wünnemann, K., Collins, G. S. & Melosh, H. J. A strain-based porosity model for use in hydrocode simulations of impacts and implications for transient crater growth in porous targets. *Icarus* **180**, 514–527 (2006).
13. Melosh, H. J. *Impact Cratering: a Geologic Process* (Oxford Univ. Press, 1989).
14. Weiss, B. P. & Elkins-Tanton, L. T. Differentiated planetesimals and the parent bodies of chondrites. *Annu. Rev. Earth Planet. Sci.* **41**, 529–560 (2013).
15. Bland, P. A. *et al.* Volatile fractionation in the early solar system and chondrule/matrix complementarity. *Proc. Natl Acad. Sci. USA* **102**, 13755–13760 (2005).
16. Asphaug, E., Jutzi, M. & Movshovitz, N. Chondrule formation during planetesimal accretion. *Earth Planet. Sci. Lett.* **308**, 369–379 (2011).
17. Sanders, I. S. & Scott, E. R. D. The origin of chondrules and chondrites: debris from low-velocity impacts between molten planetesimals? *Meteorit. Planet. Sci.* **47**, 2170–2192 (2012).
18. Taylor, G. J., Scott, E. R. D. & Keil, K. Cosmic setting for chondrule formation. In *LPI Conf. on 'Chondrules and their Origins'* **493**, abstract 58, <http://adsabs.harvard.edu/abs/1983chto.conf..262T> (1982).
19. Mastin, L. G. & Ghiorso, M. S. Adiabatic temperature changes of magma–gas mixtures during ascent and eruption. *Contrib. Mineral. Petrol.* **141**, 307–321 (2001).
20. Minton, D. A. & Levison, H. F. Planetesimal-driven migration of terrestrial planet embryos. *Icarus* **232**, 118–132 (2014).
21. Morbidelli, A., Bottke, W. F., Nesvorný, D. & Levison, H. F. Asteroids were born big. *Icarus* **204**, 558–573 (2009).
22. Connelly, J. N. *et al.* The absolute chronology and thermal processing of solids in the solar protoplanetary disk. *Science* **338**, 651–655 (2012).
23. Hood, L. L. & Weidenschilling, S. J. The planetesimal bow shock model for chondrule formation: a more quantitative assessment of the standard (fixed Jupiter) case. *Meteorit. Planet. Sci.* **47**, 1715–1727 (2012).
24. Evans, N. J. I. *et al.* The Spitzer c2d legacy results: star-formation rates and efficiencies; evolution and lifetimes. *Astrophys. J.* **181** (Suppl.), 321–350 (2009).
25. O'Brien, D. P., Morbidelli, A. & Bottke, W. F. The primordial excitation and clearing of the asteroid belt—revisited. *Icarus* **191**, 434–452 (2007).
26. Melosh, H. J. & Sonett, C. P. When worlds collide—jetted vapor plumes and the Moon's origin. In *LPI Conf. on 'Origin of the Moon'* **1**, 621–642, <http://adsabs.harvard.edu/abs/1986ormo.conf.621M> (1986).
27. Vickery, A. M. The theory of jetting: application to the origin of tektites. *Icarus* **105**, 441–453 (1993).
28. Weidenschilling, S. J. Initial sizes of planetesimals and accretion of the asteroids. *Icarus* **214**, 671–684 (2011).
29. Hezel, D. C. & Palme, H. The conditions of chondrule formation, Part I: Closed system. *Geochim. Cosmochim. Acta* **71**, 4092–4107 (2007).
30. Cuzzi, J. N. & Alexander, C. M. O. Chondrule formation in particle-rich nebular regions at least hundreds of kilometres across. *Nature* **441**, 483–485 (2006).

**Acknowledgements** We thank I. Sanders for his review, which improved the manuscript. We gratefully acknowledge the developers of iSALE ([www.isale-code.de/projects/iSALE](http://www.isale-code.de/projects/iSALE)), especially G. Collins, K. Wünnemann, D. Elbeshausen and B. Ivanov. This research was supported by NASA grant number PGG NNX10AU88G.

**Author Contributions** While working on calculating the cooling rates of impact-produced chondrules with H.J.M., B.C.J. conceived the idea that chondrules could form by jetting during low-velocity accretionary impacts. D.A.M. produced the Monte Carlo accretion code results. B.C.J. produced the hydrocode and radiative transfer code results. All authors contributed to preparation of the manuscript and the conclusions presented in this work.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.C.J. ([brjohns@mit.edu](mailto:brjohns@mit.edu)).



## METHODS

**Droplet sizes jetting model.** Assuming that a model for melt droplet formation in ejecta curtains<sup>8</sup> can be extended to droplet formation in jetted material, we calculate the size of chondrules with ejection velocities equal to the impact velocity (Extended Data Fig. 2). However, jetting ejects material at velocities above the impact velocity and higher ejection velocities yield smaller droplets. Consequently, the droplet sizes in Extended Data Fig. 2 are the maximum size of droplets produced by jetting. For chondrule-forming impacts with impact velocities of 2.5–5.5 km s<sup>-1</sup> and impactor sizes of 100–1,000 km, jetted material will create millimetre-scale droplets consistent with the observed size of chondrules (Extended Data Fig. 2).

In terrestrial impact ejecta deposits, larger, centimetre-scale tektites (or melt fragments) are found along with millimetre-scale melt droplet spherules<sup>8</sup>, whereas chondrules are all roughly the same size<sup>18</sup>. For terrestrial impacts, the initial dynamic fragmentation<sup>31</sup> of melt forms centimetre-scale droplets<sup>8</sup>. Upon release from high pressure, more highly shocked material or more-volatile-rich material separates into vapour and liquid. Acceleration of this vapour–liquid mix produces a differential velocity between droplets and the surrounding vapour and the balance of inertial forces and surface tension determines the droplet size<sup>8</sup>. The droplets formed in this accelerated two-phase mixture are roughly millimetre-scale for terrestrial impact conditions. The volatile-rich nature of chondrule precursor material may explain the apparent lack of tektite-like chondrules. Essentially, if any silicates are shock-melted there will be enough vaporized volatiles present that the material will be best described by a vapour–liquid mixture.

**Droplet sizes splashing model.** Droplet size estimates from the splashing model for chondrule formation<sup>16</sup> assume that the  $pv$  component of specific enthalpy  $h = q + pv$  (where  $q$  is specific internal energy,  $p$  is pressure, and  $v$  is specific volume) is converted to surface energy of droplets with nearly perfect efficiency. Cavitation and droplet formation does not occur until  $p < 0$  and when  $p = 0$  the  $pv$  component of enthalpy is also zero. Thus, when ex-solution of gases is ignored, isenthalpic release from pressure  $p$  causes decompression heating (that is, the  $pv$  part of enthalpy is converted to internal energy  $q$ )<sup>19</sup>. Even for a magma releasing to zero pressure from 200 MPa the associated temperature change is only 50 K (ref. 19). Because there is no  $pv$  component of enthalpy at the time of fragmentation, the assumption that all enthalpy is converted to surface energy is unphysical. The actual size of the ‘droplets’ is then determined by the balance of inertial forces and tensional forces<sup>31</sup>. For typical strain rates of  $\dot{\epsilon} \approx 1 \text{ h}^{-1}$  (ref. 16), surface tension of  $\sigma = 0.3 \text{ N m}^{-1}$ , density  $\rho = 3,000 \text{ kg m}^{-3}$  splashing will create ‘droplets’ with diameter  $d = (40\sigma/\rho\dot{\epsilon}^2)^{1/3} \approx 40 \text{ m}$  (refs 8 and 31).

**Hydrocode modelling.** Extended Data Table 1 describes the input parameters used in our iSALE models. The number of equations of state that accurately represent geologic materials under the extreme conditions occurring during impacts is small. Here we use equations of state for dunite<sup>32</sup>, produced by the ANEOS program<sup>33</sup>, to approximate the bulk properties of the impactors and target bodies during accretion. We use dunite strength, thermal softening and porous compaction parameters from refs 34–36 and references therein.

During the jetting process, material is first shocked and then adiabatically compressed<sup>7</sup>. Consequently, we cannot calculate the amount of material melted during an impact using the peak pressure. Instead, we use the post-release temperature to estimate the amount of melt. We consider any mass that releases to temperatures above 1,373 K to be potentially chondrule-forming mass. We use the solidus temperature because ANEOS does not account for the latent heat of fusion. Hence, temperatures above the solidus are exaggerated and cannot be trusted<sup>37</sup>. Additionally, ANEOS in its current form tends to underpredict the entropy of geologic material shocked to a given pressure<sup>38,39</sup>. This means that ANEOS tends to underestimate the degree of melting and the temperatures after release.

For a given impact velocity the total mass of melt ejected above some ejection velocity  $v_{ej}$  is calculated by summing the mass of Lagrangian tracers with  $v > v_{ej}$  that also release to temperatures above 1,373 K. For a more detailed description of Lagrangian tracers and how ejection velocities are determined see ref. 8. For each impactor and target porosity, we created lookup tables for the melt mass ejected at greater than escape velocity as a function of impact velocity and escape velocity. These tables cover  $v_{imp}/v_{esc} = 0.5$ –2 stepping by 0.1, and  $v_{imp} = 2.5$ –6 km s<sup>-1</sup> stepping by 0.5 km s<sup>-1</sup>. Using these tables and hydrodynamic similarity<sup>13</sup>, we estimate the amount of melt created by all of the impacts occurring in the GAME model. For impacts with  $v_{imp}/v_{esc}$  and/or  $v_{imp}$  falling between the data points from our iSALE models, we use bi-linear interpolation to estimate the fraction of impactor mass that is melted and ejected at higher than escape velocity.

To check the assumption of hydrodynamic similarity we ran a model with an impactor of 1,000 km diameter and found that the amount of melted and jetted material, normalized by impactor mass, did not change. We also made a run with the target and impactor initially having intact rock strength. This run was only minimally different, less than 1%, from the results using an originally damaged impactor and target.

Thus, we conclude that material strength has a very limited effect on jetting efficiency at least for impact velocities above 2.5 km s<sup>-1</sup>. However, as porosity increases, the overall jetting efficiency decreases significantly. This is consistent with jetting being less efficient in more compressible materials<sup>7</sup>.

**Code availability.** At present, iSALE is not fully open source. It is distributed on a case-by-case basis to academic users in the impact community, strictly for non-commercial use. Scientists interested in using or developing iSALE should see [http://www.isale-code.de/redmine/projects/isale/wiki/Terms\\_of\\_use](http://www.isale-code.de/redmine/projects/isale/wiki/Terms_of_use) for a description of application requirements. The one-dimensional radiative transfer code used here is available on request from B. Johnson (brjohns@mit.edu). GAME is available upon request from D. Minton (daminton@purdue.edu).

**Monte Carlo accretion.** The minimum mass solar nebula (MMSN) describes a disk that is just massive enough to create the observed planets. The MMSN implicitly assumes that no mass is lost from the solar nebula during the planet formation process. More massive solar nebulae are often considered by dynamical modellers<sup>20</sup>. In the 3MMSN model a small fraction of impacts have  $v_{imp} > 6 \text{ km s}^{-1}$  and  $v_{imp}/v_{esc} > 2$ . In our chondrule mass calculation, we do not include these impacts. This omission changes our mass estimate by only a few per cent, an error that is much smaller than the uncertainties inherent to our calculation.

We modelled disks with a main-belt-like size frequency distribution between 100 km and 1,000 km (ref. 21). This size frequency distribution (SFD) was proposed<sup>21</sup> because it can produce embryos of larger than the Moon’s mass before the protoplanetary disk dissipates ( $\sim 3 \text{ Myr}$ ). The early formation of embryos is required to explain the mass depletion and dynamical state of the main belt<sup>25</sup>. Other models show that an initially monodisperse population of planetesimals of 0.1 km diameter can create larger-than-lunar mass embryos and the observed size frequency of main-belt asteroids in just  $10^5 \text{ yr}$  (ref. 28). Because GAME tracks the accretion history of each object in the model, starting with planetesimals of 0.1 km diameter is computationally unfeasible. However, we did model disks with a distribution of bodies generated from a Gaussian distribution centred at 100 km in diameter with a 50-km standard deviation (truncated at 0 km).

The MMSN model with a Gaussian SFD produced 0.86, 0.86, and 1.9 times the mass of chondrules made in the MMSN model with a main-belt like SFD, for 1%, 10%, and 25% porous cases, respectively. The 3MMSN model with a Gaussian SFD produced 0.52, 0.45, and 0.22 times the mass of chondrules made in the 3MMSN model with a main-belt-like SFD, for 1%, 10%, and 25% porous cases, respectively. We also found this Gaussian SFD produced chondrules somewhat closer to the Sun. For the MMSN models, by 5 Myr the Gaussian SFD produced chondrules at 1.6 AU from the Sun, whereas the main-belt-like SFD produced chondrules at 2.2 AU. For the 3MMSN models, by 5 Myr the Gaussian SFD produced chondrules at 2.4 AU, whereas the main-belt-like SFD produced chondrules at 3.9 AU.

**Radiative transfer and cooling rates.** We find the geometry of a spherically expanding plume, as used by ref. 40, is unrealistic for jetted melt. To estimate the cooling rates of impact-produced melt droplets, we model the ejected material as an infinite sheet with a density that decreases with time. We assume this geometry so we can use a one-dimensional radiative transfer code to determine the cooling rates. Our radiative transfer code uses the diffusion approximation<sup>41</sup>. We benchmarked the code using the non-equilibrium Marshak diffusion problem<sup>42</sup>. Extended Data Fig. 3 schematically shows our assumed geometry, where

$$R_{in}(t) = R_{in}(t_0) + v_{in}t$$

$$R_{out}(t) = R_{out}(t_0) + v_{out}t$$

We consider the case where  $v_{in} = v_{imp} = 3 \text{ km s}^{-1}$  and  $v_{out} = 3.5 \text{ km s}^{-1}$ . At 2.5 s into the impact of a 10-km-diameter projectile,  $R_{in} = 8.8 \text{ km}$  and  $R_{out} = 9.5 \text{ km}$ . This region has an average density of  $\rho_{jet} = 800 \text{ kg m}^{-3}$  and an average thickness of  $h = 215 \text{ m}$ . The  $7 \times 10^{12} \text{ kg}$  of material represents 48% of the total melt mass ejected with velocities above 3 km s<sup>-1</sup>. This is the result of lower-velocity material dominating the mass of the jet<sup>7</sup>. When considering larger impactor sizes we use hydrodynamic scaling to produce different initial conditions.

We focus on the mass-averaged cooling rate of this material as an estimate for the average cooling rate of chondrules created by jetting. If we were to consider faster parts of the plume, we would expect higher cooling rates; slower ejecta would have lower cooling rates.

We assume the heat capacity of the droplets to be  $1,000 \text{ J kg}^{-1} \text{ K}^{-1}$  and the bulk density of the droplets is  $\rho_{drop} = 3,000 \text{ kg m}^{-3}$ . The droplets are assumed to be black-bodies and thus have a collective opacity of  $\kappa = 3\phi/4r_{drop}$ , where  $\phi$  is the fraction of the volume occupied by the droplets and  $r_{drop}$  is the radius of the of the melt droplets. For simplicity, we neglect the opacity of any vapour that may be present, that is, impact-produced vapour or gas in the solar nebula. We assume that the droplets start out with temperatures of 2,000 K. We also set the constant temperature boundary condition to 300 K.

The opacity is updated using the volume calculated at each time step:

$$V(t) = \pi(R_{\text{out}}^2 - R_{\text{in}}^2)dh$$

where  $dh = h/800$  is the thickness of one of 400 equal-sized computational cells. Note that for the reflective boundary condition and symmetry of the problem we model only half of the thickness of the jet. Then the volume fraction occupied by molten droplets is:

$$\phi = \frac{M_0}{\rho_{\text{drop}} V}$$

where

$$M_0 = \rho_{\text{jet}} V(t_0)$$

In our approximation of the geometry of a jet, radiation escapes only from the free surface of the jet. We find for impactors 100–1,000 km in size that the cooling of the jet takes hours to days (Fig. 4). The fastest part of the jet moves a few kilometres per second faster than the parts that only just escape the target body. Thus, the jet has a horizontal extent of  $10^4$ – $10^5$  km on the timescale of cooling (estimated by multiplying a few kilometres per second by an hour to a day). The distance over which radiation must diffuse is much larger than the thickness scale, which is 2–20 km for impactors that are 100–1,000 km in diameter. Additionally, the free surface is much colder (here assumed to be 300 K) when compared to the temperature of adjacent material along the jet. Thus, our assumption that radiative transfer along the jet is negligible is quite reasonable.

Extended Data Fig. 4 shows the results of our radiative transfer code, assuming a droplet diameter of 1 mm and an impactor diameter of 1,000 km. Extended Data Fig. 4 shows that the maximum mass-averaged cooling rate, which is plotted in Fig. 4, occurs when the mass-averaged temperature is  $\sim 1,400$  K. The plots also show that the outer part of the jet cools at about ten times the rate of the inner part of the jet. The outer part of the jet is not to be confused with the faster-moving part of the jet (see Extended Data Fig. 3), although faster-moving parts of the jet probably have higher cooling rates too.

**Dust enrichment and chondrule–chondrule collisions.** Turbulent mixing of nebular gas and jetted material lead to chondrule–chondrule collisions. In the jet the turbulent velocity is approximately a few per cent of the flow velocity, which is the product of the jet thickness and local velocity gradient<sup>8</sup>. This velocity is independent of impactor size and our iSALE models indicate that the turbulent velocity of the jet is  $v_{\text{turb}} \approx 5 \text{ m s}^{-1}$ . However, the turbulent velocity may be a few per cent of the ejection velocity,  $v_{\text{turb}} \approx 100 \text{ m s}^{-1}$ , if turbulence created at the jet–nebula interface is dominant. The relative velocities of inertial particles in turbulent two-fluid flow is<sup>42</sup>:

$$v_{\text{rel}} = v_{\text{turb}} \left( 1 + 1.5 \tau_p \frac{v_{\text{turb}}}{l_{\text{turb}}} \right)^{-1/2}$$

where  $l_{\text{turb}}$  is the turbulent length scale taken to be approximately equal to the jet thickness of 2.2–22 km for impactor diameters of 100–1,000 km. In the midplane at around 3 AU,  $\tau_p = 4.9 \times 10^3 \text{ s}$  for a 1-mm-diameter particle, with longer times occurring outside the midplane<sup>43</sup>. This equation is valid for inertial or heavy particles that have  $\tau_p \geq l_{\text{turb}}/v_{\text{turb}}$  and when the collisional mean free path is larger than the correlation length<sup>44</sup>. In a turbulent flow, the correlation of velocity along a line connecting two points separated by  $\lambda$  is<sup>42</sup>:

$$f = \max \left( 0, 1 - \frac{0.9 \epsilon^{2/3} \lambda^{2/3}}{v_{\text{turb}}^2} \right)$$

where dissipation rate  $\epsilon \approx v_{\text{turb}}^3/l_{\text{turb}}$ . When  $f = 0$  the velocities are uncorrelated;  $f = 1$  means that the velocities at the two points are identical. The minimum value of  $\lambda$  that yields  $f = 1$  is the correlation length. We find that at early times the collisional mean free path  $\lambda$  is much smaller than the correlation length. We therefore introduce the following term to account for particles having initially partially correlated velocities:

$$v_{\text{rel}} = v_{\text{turb}} \left( 1 + 1.5 \tau_p \frac{v_{\text{turb}}}{l_{\text{turb}}} \right)^{-1/2} (1 - f)$$

Assuming a collision efficiency of unity, the average number of collisions a single chondrule experiences per unit time is<sup>42</sup>:

$$N \approx 4\pi^{1/2} d^2 v_{\text{rel}} n_c$$

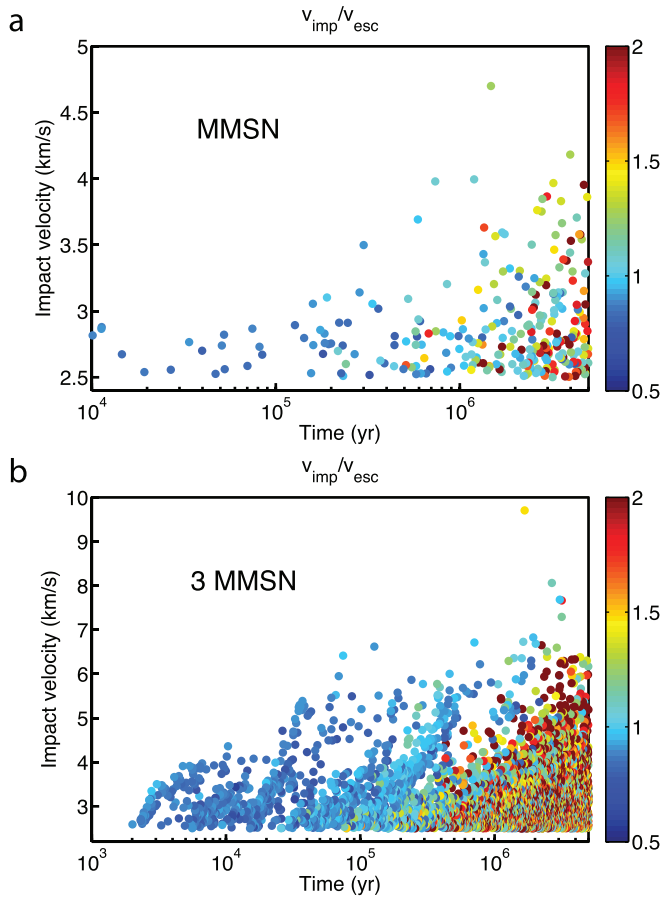
so that the mean free path  $\lambda = (4\pi^{1/2} d^2 n_c)^{-1}$ . Extended Data Fig. 4 shows the results for number density, relative velocity, rate of collisions and cumulative number of collisions calculated at one-second intervals and using the bulk density coming from the radiative transfer sections. We note that the turbulent velocity is assumed to be constant, although it probably decays with time.

While the drops are still above their solidus, the collision velocities shown in Extended Data Fig. 4b are low and could result in coalescence or bouncing<sup>43</sup>. As time goes on the relative velocities increase, but partially molten, cooling chondrules can survive impacts at velocities up to  $100 \text{ m s}^{-1}$  (ref. 43). Although our simple calculations show that individual chondrules may experience many collisions (Extended Data Fig. 4c, d), it is unknown what fraction of chondrule collisions will result in the formation of compound chondrules, and the fraction of compound chondrules provides only a minimum estimate of the chondrule concentrations<sup>44,45</sup>.

Because we do not include nebular gas in our hydrocode models, the pressure in the jet drops below typical nebular pressures of  $\sim 0.01$ – $100 \text{ Pa}$  (ref. 3). However, even at background nebula pressures the dust enrichment of the jetted material is enough to explain the volatile content of chondrule olivine. If chondrules formed at typical nebular pressures, dust enrichments greater than  $10^6$  (where dust enrichment is the solid-to-vapour ratio relative to a system of solar composition<sup>3–5</sup>) are needed to explain the volatile content of chondrule olivine<sup>3–5</sup>. For 1-mm-diameter chondrules with densities of  $3,000 \text{ kg m}^{-3}$  this corresponds to number densities higher than  $10^3$ – $10^4 \text{ m}^{-3}$  depending on the total pressure<sup>4</sup>. For the geometry of the jet described in the radiative transfer section and for a 100-km-diameter impactor, the number density will remain above  $10^4 \text{ m}^{-3}$  until 1.2 h after ejection and drops below  $10^3 \text{ m}^{-3}$  after 3.9 h (Extended Data Fig. 4a). This time scales linearly with impactor size, meaning that the bulk density of material ejected by a 1,000-km-diameter body remains above  $10^4 \text{ m}^{-3}$  for about 12 h and above  $10^3 \text{ m}^{-3}$  for 39 h. A comparison of Extended Data Fig. 4 and Fig. 4 shows that this is more than enough time for chondrules to cool below the solidus at  $\sim 1,400 \text{ K}$ . Our model also predicts that chondrules that experience higher cooling rates (Extended Data Figs 4 and 5) will cool to the solidus at higher dust enrichments than those made by the same impact with lower cooling rates.

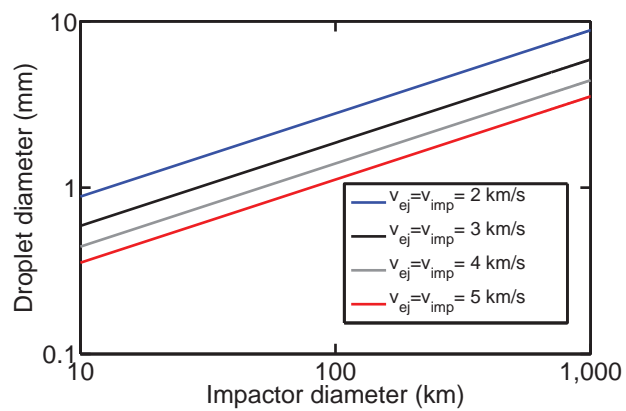
Both dust enrichment and total vapour pressure are important for determining the volatile content of chondrule olivine<sup>3</sup>. To obtain more robust estimates of the time history of dust enrichment, the rate of chondrule–chondrule collisions, and the total vapour pressure, the dynamic interaction of jetted material with nebular gas is required. This requires a two-fluid hydrodynamic code that can accurately model the interactions of particles and gas.

1. Benz, W., Cameron, A. & Melosh, H. J. The origin of the Moon and the single-impact hypothesis III. *Icarus* **81**, 113–131 (1989).
2. Thompson, S. L. & Lauson, H. S. Improvement in the Chart D Radiation-Hydrodynamic Code. III. Revised analytic equations of state. *Sandia Report* SC-RR-71 0174 (1984).
3. Grady, D. E. Local inertial effects in dynamic fragmentation. *J. Appl. Phys.* **53**, 322 (1982).
4. Potter, R. W. K., Collins, G. S., Kiefer, W. S., McGovern, P. J. & Kring, D. A. Constraining the size of the South Pole-Aitken basin impact. *Icarus* **220**, 730–743 (2012).
5. Davison, T. M., Collins, G. S. & Ciesla, F. J. Numerical modelling of heating in porous planetesimal collisions. *Icarus* **208**, 468–481 (2010).
6. Bowling, T. J. et al. Antipodal terrains created by the Rheasilvia basin forming impact on asteroid 4 Vesta. *J. Geophys. Res. Planets* (2013).
7. Collins, G. S. & Melosh, H. J. Improvements to ANEOS for multiple phase transitions. *Lunar Planet. Sci. Conf.* **266A** (2014).
8. Kraus, R. G. et al. Shock vaporization of silica and the thermodynamics of planetary impact events. *J. Geophys. Res.* **117**, E09009 (2012).
9. Kurosawa, K. et al. Shock-induced silicate vaporization: the role of electrons. *J. Geophys. Res.* **117**, E04007 (2012).
10. Johnson, B. C., Minton, D. A. & Melosh, H. J. The impact origin of chondrules. *Lunar Planet. Sci. Conf.* **1471** (2014).
11. Bingjing, S. & Olson, G. L. Benchmark results for the non-equilibrium Marshak diffusion problem. *J. Quant. Spectrosc. Radiat. Transf.* **56**, 337–351 (1996).
12. Abrahamson, J. Collision rates of small particles in a vigorously turbulent fluid. *Chem. Eng. Sci.* **30**, 1371–1379 (1975).
13. Weidenschilling, S. J. Particles in the nebular midplane: collective effects and relative velocities. *Meteorit. Planet. Sci.* **45**, 276–288 (2010).
14. Ormel, C. W. & Cuzzi, J. N. Closed-form expressions for particle relative velocities induced by turbulence. *Astron. Astrophys.* **466**, 413–420 (2007).
15. Ciesla, F. J. Chondrule collisions in shock waves. *Meteorit. Planet. Sci.* **41**, 1347–1359 (2006).
16. Wünnemann, K., Collins, G. S. & Osinski, G. R. Numerical modelling of impact melt production in porous rocks. *Earth Planet. Sci. Lett.* **269**, 530–539 (2008).
17. Collins, G. S., Melosh, H. J. & Wünnemann, K. Improvements to the  $\epsilon$ - $\alpha$  porous compaction model for simulating impacts into high-porosity solar system objects. *Int. J. Impact Eng.* **38**, 434–439 (2011).

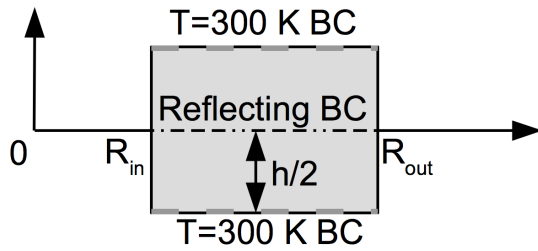


**Extended Data Figure 1 | Timing and velocity of chondrule-forming impact.** Chondrule-forming impacts with velocities above  $2.5 \text{ km s}^{-1}$  for the MMSN model (a) and the 3MMSN model (b). The points are coloured according to  $v_{\text{imp}}/v_{\text{esc}}$  (shown on the colour scale). Note that  $v_{\text{imp}}/v_{\text{esc}}$  may be less than one because  $v_{\text{esc}}$  is considered to be the escape velocity after the target and impactor have combined to form a more massive body.

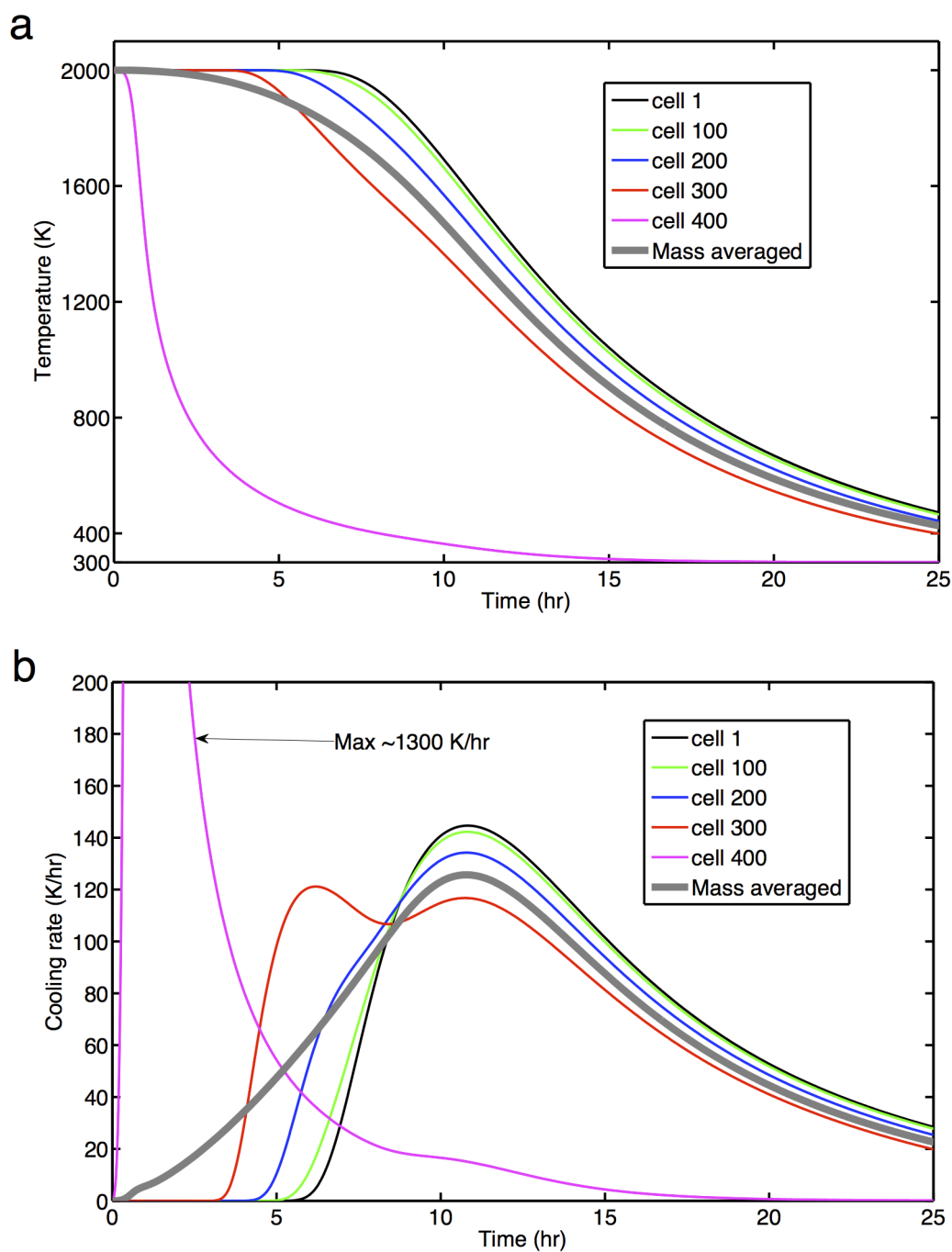




**Extended Data Figure 2 | Maximum size of droplets created by jetting.** The different lines represent different impact velocities, as indicated.



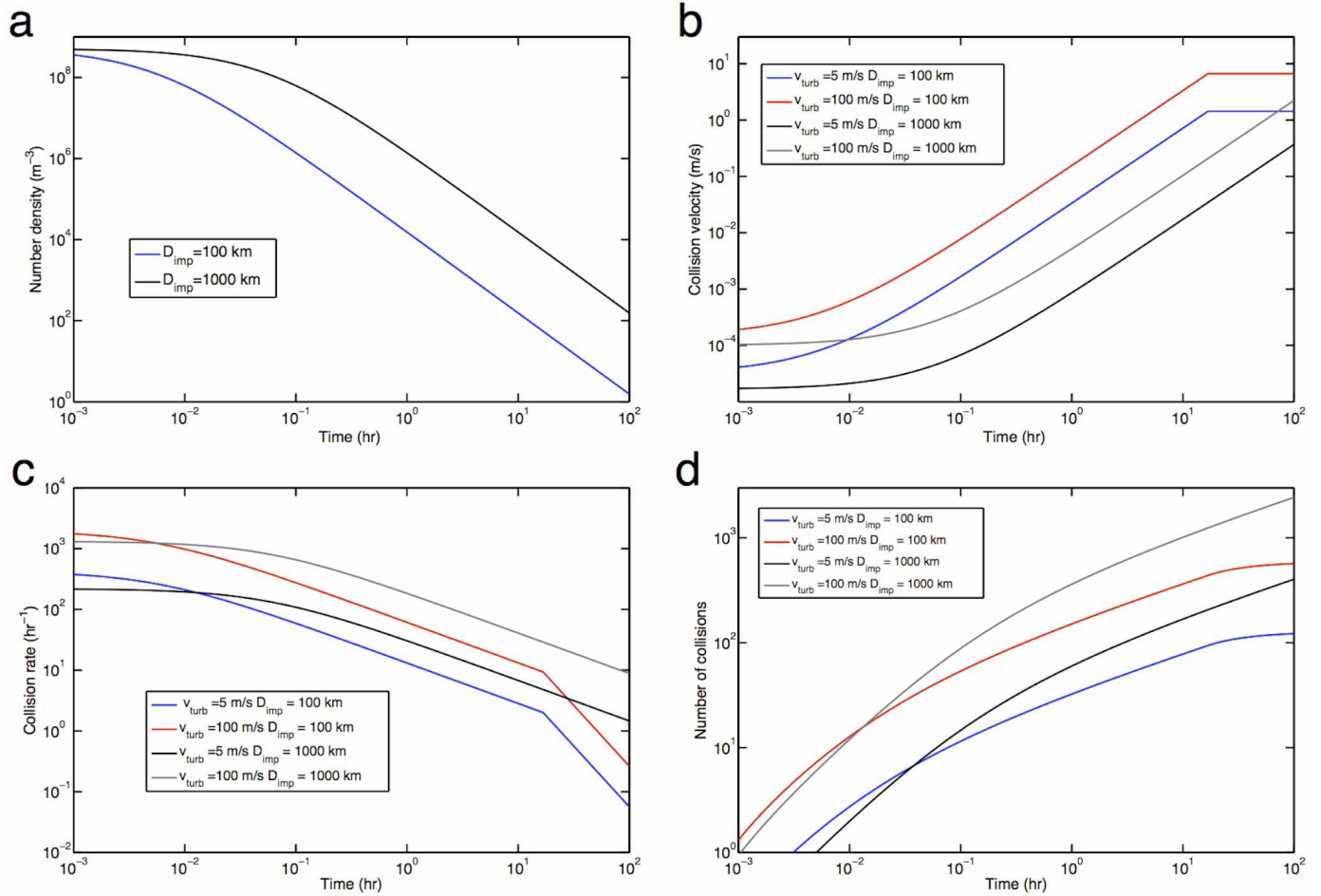
**Extended Data Figure 3 | Schematic showing the geometry of our radiative transfer models.** The horizontal axis shows radial distance from the point of impact. The vertical axis marks the thickness of the jet. We model a portion of the jet as an annulus that moves outward radially. The width of this annulus also grows with time. BC, boundary condition;  $h$ , the thickness of the jet.



**Extended Data Figure 4 | Temperature time history for a jet consisting of 1-mm-diameter droplets created by a 1,000-km-diameter impactor.** The different coloured curves represent different computational cells, as indicated, where cell 1 is the innermost cell, which has a reflective boundary condition on one side, and cell 400 is the outermost cell, which radiates into a background

at 300 K. The thick grey curve is the mass-averaged temperature, which we use as proxy for the average temperature of material in the plume. Panel **a** shows the temperature as a function of time, while **b** shows the cooling rate as a function of time.





**Extended Data Figure 5 | Chondrule density and collision rates.**

**a**, The number density of chondrules is plotted as a function of time for 100-km-diameter and 1,000-km-diameter impactors. **b**, Relative collision

velocity plotted as a function of time. **c**, Rate of collisions a single chondrule experiences plotted as a function of time. **d**, Cumulative number of impacts a chondrule experiences plotted as a function of time.

Extended Data Table 1 | iSALE input parameters

Description	Input
Equation of state	ANEOS dunite <sup>a</sup>
Melting temperature <sup>b</sup>	1373 K
Thermal softening parameter <sup>b</sup>	1.1
Simon A parameter <sup>b</sup>	1520 MPa
Simon B parameter <sup>b</sup>	4.05
Poisson's ratio $\nu$	0.25
Frictional coefficient (damaged) $\mu^c$	0.63
Frictional coefficient (undamaged) $\mu^c$	1.58
Strength at infinite pressure $Y_m^c$	3.26 GPa
Cohesion (damaged) $Y_0^c$	10 kPa
Cohesion (undamaged) $Y_0^c$	5.07 MPa
Strain at which porous compaction begins $\varepsilon_e^d$	0.01
Rate of porous compaction $\kappa^d$	0.98
Size of high resolution cell	12.5 m
Number of high resolution cells horizontal direction	1500
Number of high resolution cells vertical direction	1600
Target and projectile temperature	300 K

<sup>a</sup> See ref. 31.<sup>b</sup> See ref. 46 and references therein for a description of Simon parameters, the thermal softening parameter, and their implementation in iSALE.<sup>c</sup> See ref. 11 and references therein for a description of strength model parameters and their implementation in iSALE.<sup>d</sup> See refs 12 and 47 for a description of the porous compaction model parameters and their implementation in iSALE.

# Direct observation of electron propagation and dielectric screening on the atomic length scale

S. Neppel<sup>1,2,†</sup>, R. Ernstorfer<sup>3</sup>, A. L. Cavalieri<sup>4,5,6</sup>, C. Lemell<sup>7</sup>, G. Wachter<sup>7</sup>, E. Magerl<sup>2</sup>, E. M. Bothschafter<sup>8</sup>, M. Jobst<sup>1,2</sup>, M. Hofstetter<sup>2,8</sup>, U. Kleineberg<sup>2,8</sup>, J. V. Barth<sup>1</sup>, D. Menzel<sup>1,3</sup>, J. Burgdörfer<sup>7,9</sup>, P. Feulner<sup>1</sup>, F. Krausz<sup>2,8</sup> & R. Kienberger<sup>1,2</sup>

The propagation and transport of electrons in crystals is a fundamental process pertaining to the functioning of most electronic devices. Microscopic theories describe this phenomenon as being based on the motion of Bloch wave packets<sup>1</sup>. These wave packets are superpositions of individual Bloch states with the group velocity determined by the dispersion of the electronic band structure near the central wavevector in momentum space<sup>1</sup>. This concept has been verified experimentally in artificial superlattices by the observation of Bloch oscillations<sup>2</sup>—periodic oscillations of electrons in real and momentum space. Here we present a direct observation of electron wave packet motion in a real-space and real-time experiment, on length and time scales shorter than the Bloch oscillation amplitude and period. We show that attosecond metrology<sup>3</sup> (1 as = 10<sup>−18</sup> seconds) now enables quantitative insight into weakly disturbed electron wave packet propagation on the atomic length scale without being hampered by scattering effects, which inevitably occur over macroscopic propagation length scales. We use sub-femtosecond (less than 10<sup>−15</sup> seconds) extreme-ultraviolet light pulses<sup>4</sup> to launch photoelectron wave packets inside a tungsten crystal that is covered by magnesium films of varied, well-defined thicknesses of a few ångströms<sup>1</sup>. Probing the moment of arrival of the wave packets at the surface with attosecond precision reveals free-electron-like, ballistic propagation behaviour inside the magnesium adlayer—constituting the semi-classical limit of Bloch wave packet motion. Real-time access to electron transport through atomic layers and interfaces promises unprecedented insight into phenomena that may enable the scaling of electronic and photonic circuits to atomic dimensions. In addition, this experiment allows us to determine the penetration depth of electrical fields at optical frequencies at solid interfaces on the atomic scale.

A detailed microscopic understanding and control of electronic and optical properties of solids depends on our ability to access the dynamics of electrons on atomic time and length scales. Tracking the propagation of electrons in real time requires the ability to pinpoint their position at a rate comparable to the time on which interactions with other electrons and the crystal lattice may affect their trajectories inside the material. In a classical picture, the upper bound for the necessary temporal resolution is therefore the time it takes the electrons to travel the several-ångström distance between neighbouring atoms. This implies transit times well below one femtosecond even for kinetic energies as low as  $E_{\text{kin}} \approx 1$  eV. Previous time-resolving studies on electron propagation in condensed matter employed laser-based spectroscopic techniques to reveal ballistic currents and drift motion of charge carriers<sup>5–8</sup>. Restricted to the pico- and femtosecond timescales, they are able to probe carrier dynamics averaged over several hundreds of nanometres. In contrast, the experimental approach demonstrated here offers

quantitative real-time access to electron transport on the inter-atomic length scale.

Figure 1 illustrates the basic principle of the experiment. Attosecond extreme-ultraviolet (XUV) light pulses generate Bloch electron wave packets with final-state energies substantially above the vacuum level. Electron wave packets with a sufficiently large momentum component along the surface-normal direction  $z$  contribute to the photoelectron current reaching the time-of-flight detector. Conventional photoelectron spectroscopy is restricted to measuring the energy and momentum distributions of photoelectrons. Here we additionally capture the temporal profile of the photoemission process by having the ejected electrons interact with the controlled few-cycle electric field of a visible/near-infrared (henceforth referred to as ‘NIR’) laser pulse covering the 500–1,000 nm spectral range and synchronized, with attosecond precision, to the XUV pulse.

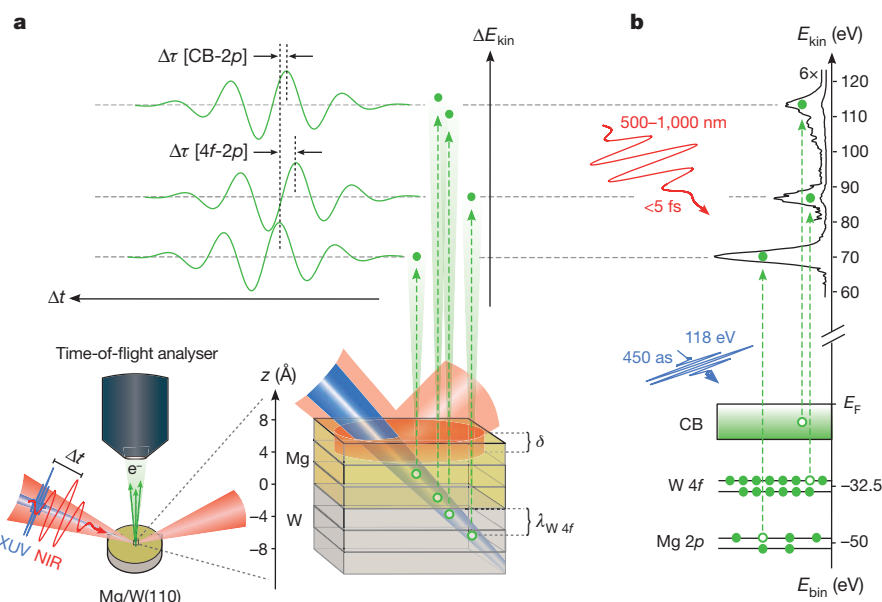
As this field modifies (‘streaks’) the momentum of a photoelectron in proportion to the laser vector potential  $A_L(z, t)$  at the instant  $t$  the electron enters the NIR field<sup>13,9,10</sup>, the temporal profile of the electron wave packet leaving the sample is mapped onto its final momentum distribution. Full streaking spectrograms obtained by recording these laser-modulated electron energy distributions as a function of delay  $\Delta t$  between the XUV and the NIR pulses are therefore highly sensitive to the spatio-temporal characteristics of both the photoelectron wave packet and the streaking laser field on the ångström length scale and the attosecond timescale<sup>9–11</sup>.

Direct (time-domain) access to these electronic and optical wave packets promises a unique insight into the photoelectric effect, including underlying electron propagation and phenomena as fundamental as dielectric screening of light fields at solid surfaces. Here we show that this can be achieved by combining state-of-the-art attosecond timing metrology (chronoscopy)<sup>3,9–11</sup> with sample engineering on the ångström level<sup>4,12</sup>.

When excited from Bloch states inside the crystal to positive-energy states<sup>13</sup>, photoelectrons are not immediately exposed to the streaking field because the electric field amplitude is substantially reduced (or screened) at the topmost layer by the response of the metal electrons. Therefore, the time delay associated with the propagation of the respective Bloch wave packets towards the surface (included in both the quantum-mechanical one-step<sup>13</sup> and the semi-classical three-step<sup>14,15</sup> description of photoemission) is encoded in the streaking spectrogram<sup>9–11</sup>. Differences in the propagation times of electrons ejected from different initial states manifest themselves as a temporal offset between the respective streaking traces<sup>9–11</sup> (see Fig. 1a). Previous studies on single crystals revealed a considerable time delay between the emission of core-level and conduction band (CB) photoelectrons from the transition metal tungsten W(110)<sup>9</sup>, whereas such a delay was found to be absent

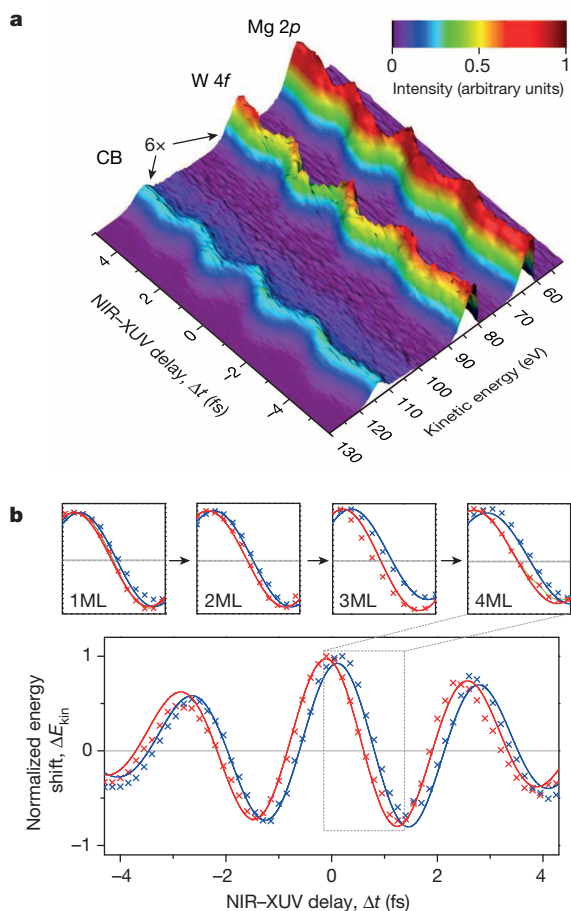
<sup>1</sup>Physik-Department, Technische Universität München, 85747 Garching, Germany. <sup>2</sup>Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Straße 1, 85748 Garching, Germany. <sup>3</sup>Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4–6, 14195 Berlin, Germany. <sup>4</sup>Max Planck Institute for the Structure and Dynamics of Matter, Luruper Chaussee 149, 22761 Hamburg, Germany. <sup>5</sup>Fakultät für Mathematik, Informatik und Naturwissenschaften, University of Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany. <sup>6</sup>Center for Free-Electron Laser Science (CFEL), Luruper Chaussee 149, 22761 Hamburg, Germany. <sup>7</sup>Institute for Theoretical Physics, Vienna University of Technology, Wiedner Hauptstrasse 8-10/E136, A-1040 Vienna, Austria. <sup>8</sup>Fakultät für Physik, Ludwig-Maximilians-Universität München, Am Coulombwall 1, D-85748 Garching, Germany. <sup>9</sup>Institute of Nuclear Research of the Hungarian Academy of Sciences (ATOMKI), 4001 Debrecen, Hungary. <sup>†</sup>Present address: Lawrence Berkeley National Laboratory, Chemical Sciences Division, Berkeley, California 94720, USA.





**Figure 1 | Spatio-temporal dynamics in attosecond photoemission from Mg/W(110).** **a**, Principle of the experiment: photoelectrons (green dots) are launched inside a tungsten W(110) crystal and a magnesium (Mg) overlayer a few ångströms thick by an XUV pulse of  $\sim 450$  as, and are detected in ultrahigh vacuum with a time-of-flight analyser. At the surface, the arrival times of electrons released from different initial states are probed by streaking their associated electron energy distributions with a  $2 \times 10^{11} \text{ W cm}^{-2}$  strong electric field delivered by a sub-5 fs broadband linearly polarized visible/near-infrared laser pulse. Relative time delays  $\Delta\tau$  developing during the propagation of the photoelectrons to the metal–vacuum interface are detected as temporal shifts between their streaked energy distributions. The time shifts  $\Delta\tau$  are

sensitive to the atomic-scale electron transport characteristics (quantified by the inelastic mean free path  $\lambda_i$ ; indicated only for the W 4f electrons), the Mg overlayer thickness and the screening behaviour of the laser field at the solid–vacuum interface. **b**, Schematic energy-level diagram for the probed electronic transitions. The central XUV photon energy of  $\sim 118$  eV allows the simultaneous excitation of Mg 2p, W 4f and the joint CB states (binding energy  $E_{\text{bin}}$  as indicated). A background-corrected photoelectron spectrum of  $n = 4$  adlayers of Mg on W(110) in the absence of the laser field is shown as the black solid line. For better visibility, the strength of the CB and W 4f signals are magnified by a factor of six.



in the photoemission from the free-electron metal magnesium Mg(0001)<sup>10</sup>. Theoretical models have addressed different contributions such as may arise from the band structure of the material<sup>9,16,17</sup>, the spatial characteristics of the initial-state wavefunctions<sup>18–21</sup>, and elastic and inelastic scattering effects<sup>9,22</sup>. These models also differ from each other in the way that the screening of the laser field at the surface is taken into account when calculating the photoemission time delays<sup>19,22,23</sup>. To isolate the atomic-scale electron propagation process from this multitude of disparate effects, we investigate hybrid metallic samples consisting of a controllable number  $n$  of Mg adlayers on a W(110) crystal<sup>12,24</sup> (see Fig. 1a and Supplementary Information for details) and contrast the measured time shifts with electron transport calculations.

In our experiments, XUV pulses with a duration of about 450 as and a photon energy of  $\hbar\omega_{\text{XUV}} = 118$  eV simultaneously generate photoelectrons from core states of the substrate (W 4f) and adlayer (Mg 2p), as well as from the energetically overlapping CB states of both materials (Fig. 1b). A representative streaking spectrogram for  $n = 4$  Mg adlayers on W(110) is shown in Fig. 2a. Despite the  $\sim 80\%$  attenuation of the W

**Figure 2 | Attosecond time-resolved photoemission from Mg/W(110).**

**a**, Representative streaking spectrogram for  $n = 4$  Mg monolayers (ML). All photoelectron spectra are corrected for the inelastic electron background signal. The strength of the CB and W 4f signals is magnified by a factor of 6 for better visibility. **b**, Exemplary timing analysis of the Mg 2p and W 4f core-level electrons: the first moments calculated from their respective kinetic energy distributions are shown as red crosses (Mg 2p) and blue crosses (W 4f) as functions of NIR–XUV delay  $\Delta t$ . A global fit of the resultant streaking traces to a parameterized waveform for the NIR vector potential (solid lines) reveals a relative time shift  $\Delta\tau[4f - 2p]$ , which can be identified with the time delay occurring during the release of the electrons from the metal surface. Insets illustrate the evolution of  $\Delta\tau[4f - 2p]$  for  $0 < n \leq 4$ . Regions exhibiting the largest gradient of the streaking field (corresponding to the highest temporal resolution) are highlighted. An analogous evaluation of  $\Delta\tau[\text{CB} - 2p]$  is presented in the Supplementary Information.

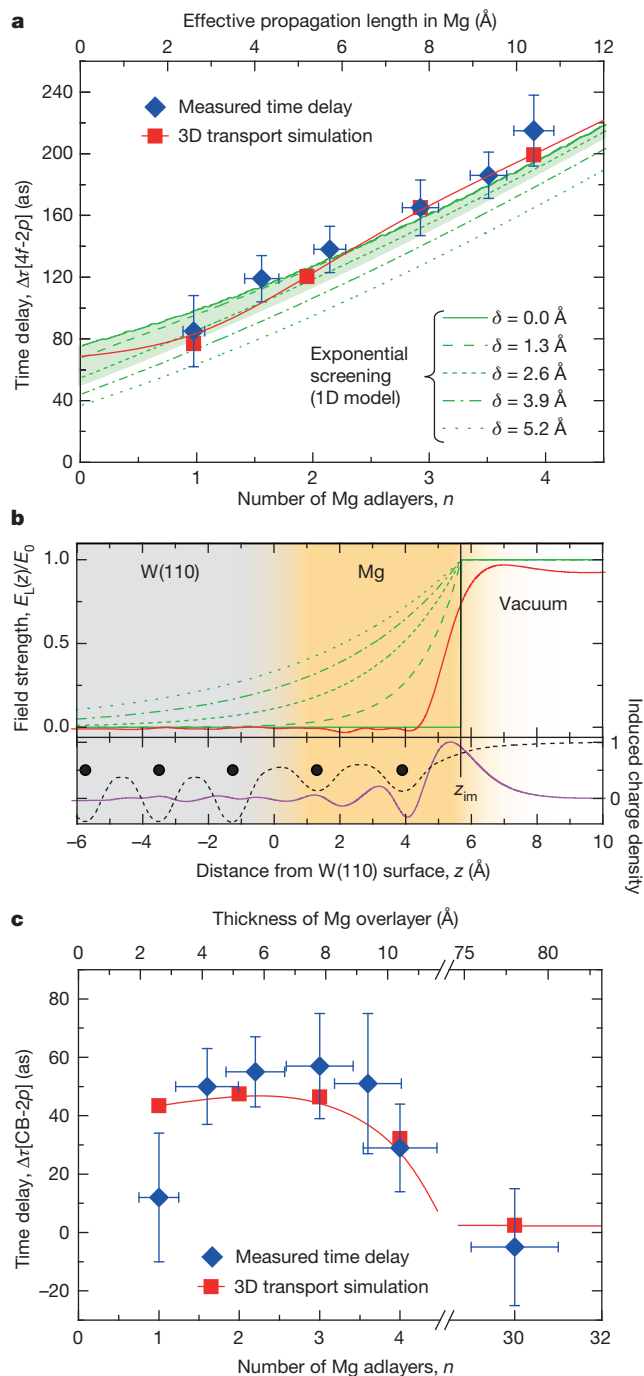
substrate photoemission due to inelastic scattering in the Mg overlayer, the streaked W 4*f* and CB photoemission lines are clearly discernible. They are also sufficiently separated from each other and from the Mg 2*p* line over the entire range of XUV–NIR delays,  $\Delta t$ , to guarantee an accurate quantitative analysis of their relative emission dynamics<sup>9–11</sup>. In what follows, we reference the emission times of the W 4*f* and CB electrons to the Mg 2*p* emission from the Mg overlayer and denote the resultant relative delays as  $\Delta\tau[4f-2p]$  and  $\Delta\tau[CB-2p]$ , respectively.

We begin with an analysis of  $\Delta\tau[4f-2p]$  because the involved photoelectrons originate from atomic-like states that are entirely localized in the W(110) substrate or the Mg overlayer. This allows unambiguous interrelation of the measured time shifts and the well-defined propagation distances in the Mg adlayer systems, which may not perfectly apply to CB electrons owing to the delocalized character of the CB initial-state

wavefunctions<sup>19–21</sup>. Relative time shifts  $\Delta\tau[4f-2p]$  extracted with a robust quantum-mechanical fitting scheme<sup>9–11</sup> (see Supplementary Information) from streaking spectrograms at different Mg coverages are summarized in Fig. 3a (blue diamonds). They reveal a distinct monotonic increase of  $\Delta\tau[4f-2p]$  with the number *n* of Mg adlayers, reaching  $\Delta\tau[4f-2p] = 215 \pm 20$  as for a Mg film 10.4 Å (*n* = 4) thick. We verified this trend using a simpler analysis that compares the first moments of the streaked W 4*f* and Mg 2*p* energy distributions as functions of  $\Delta t$  (Fig. 2b; see Supplementary Information).

The simplest description of the electron propagation is to consider ballistic motion of the centroid of a Bloch wave packet as a free point-like electron in one dimension<sup>25</sup>. For the Mg layers,  $\Delta\tau[4f-2p]$  is then dominated by the average propagation time  $\tau_{4f}$  of the 4*f*-derived wave packets with central wavevectors  $k = k_0$  travelling at a group velocity  $v_{4f} = \frac{dE(k)}{hdk}|_{k=k_0}$ . Owing to the free-electron-like band-structure of Mg<sup>10,26</sup>, we can assume that  $v_{4f} \approx \sqrt{2E_{kin}/m_e}$ , where  $m_e$  is the free electron mass. The kinetic energy of the photoelectrons inside the Mg layer amounts to  $E_{kin} \approx 93$  eV (the photon energy is  $h\omega_{XUV} = 118$  eV, the binding energy of the 4*f* electrons is  $E_b \approx 32.5$  eV, the Fermi energy of bulk Mg is  $E_{Fermi} \approx 7$  eV), leading to an average group velocity of  $v_{4f} \approx 0.057$  Å as<sup>−1</sup>. We therefore expect  $\tau_{4f}$ —which is the upper limit for  $\Delta\tau[4f-2p]$ —to increase almost linearly with the number of Mg adlayers *n*, according to  $\tau_{4f} \approx n \times d/v_{4f} \approx n \times 45$  as, where  $d = 2.6$  Å is the interlayer spacing of the epitaxial Mg films<sup>12</sup>. A linear fit to the experimental data of Fig. 3a yields a delay of  $\sim 42$  as per adlayer. The good agreement between experiment and model prediction provides conclusive evidence for the atomic-scale ballistic propagation of the 4*f* electrons being the microscopic origin of the observed time shifts in the spectrograms and also corroborates free-electron-like transport in Mg.

This interpretation is substantiated by electron transport simulations<sup>22</sup> of the ballistic motion of the W 4*f*, Mg 2*p* and CB electrons in the Mg/W(110) systems. In general, time delays obtained from such transport calculations are sensitive to (1) the average group velocities of the electrons at the relevant energies, (2) their energy-dependent inelastic mean free path  $\lambda(E_{kin})$  in the traversed materials and (3) the spatio-temporal profile of the streaking field near the surface<sup>22</sup>. The average group velocities can be deduced from electronic structure calculations<sup>9,17</sup> and all relevant values of  $\lambda(E_{kin})$  are known from synchrotron experiments<sup>10</sup>



**Figure 3 | Atomic-scale photoelectron transport and screening of the incident light field.** **a**, Time delays  $\Delta\tau[4f-2p]$  between the release of W 4*f* and Mg 2*p* electrons extracted from a large set of streaking spectrograms with different numbers of Mg adlayers are shown as blue diamonds. Error bars denote full standard deviations and are obtained by averaging measurements performed under similar experimental conditions. Fractional adlayers correspond to dispersed two-dimensional islands (on top of a completed Mg layer) that coalesce upon further Mg deposition. Green lines are time delays predicted by our one-dimensional (1D) simulation of the photoelectron release dynamics for different screening lengths  $\delta$  assuming an exponentially decaying normal component of the incident NIR streaking field at the metal–vacuum interface according to  $E_L(z, t) = E_0 e^{-(z-z_{im})/\delta}$ . The light-green-shaded area highlights the screening scenarios compatible with the experiment. Red squares indicate time delays derived from a full three-dimensional electron transport model (the red line is a guide to the eye). **b**, The upper panel illustrates the different screening scenarios for  $E_L(z, t)$  considered in **a** for the example of *n* = 2 Mg adlayers (orange shading) on W(110) (grey shading). The red line is the spatial variation of  $E_L(z, t)$  at the interface predicted by TDDFT. The lower panel is a snapshot of the NIR-induced charge density at the metal–vacuum interface at the maximum of the laser pulse derived by TDDFT (magenta line). The position of the dynamic image plane  $z_{im}$  is indicated as a vertical black solid line. The lattice potential (averaged parallel to the crystal surface) employed in the DFT calculations is shown as a dotted black line. The positions of the Mg and W atoms at the interface along the surface normal are indicated as black dots. The gradual transitions from the W electronic structure (grey) to the Mg electronic structure (orange) and to the vacuum (white) are indicated in the background colours. **c**, Comparison of time shifts  $\Delta\tau[CB-2p]$  measured between the emission of CB and Mg 2*p* electrons with time delays predicted by the three-dimensional (3D) electron transport model.

or theory<sup>27</sup>. As a consequence, our experiments open up the opportunity to explore the spatial variation of the laser field's  $E_L(z)$  component normal to the surface ( $z$  axis) on the ångström length scale.

Screening at metallic surfaces becomes effective near the so-called image plane  $z_{\text{im}}$ , located about a half-layer spacing outside the centre of the topmost atomic layer<sup>28,29</sup>. As our method probes the optical near-field at the metal–vacuum interface, the commonly used Fresnel equations based on macroscopic properties of target components with perfectly sharp interfaces cannot be applied. Instead, a phenomenological exponential decay  $E_L(z, t) = E_0 e^{-(z-z_{\text{im}})/\delta}$  of the surface-normal component of the field inside the material appears to be a reasonable assumption<sup>18,23</sup>. We therefore modelled the impact of different screening lengths  $\delta$ , that is, the length scale on which the stepwise prediction of the Fresnel formula does not apply, on the time delays  $\Delta\tau[4f-2p]$  using one-dimensional electron transport simulations (see Supplementary Information). The time delays  $\Delta\tau[4f-2p]$  predicted by this simple model as a function of  $n$  and  $\delta$  are plotted as green lines in Fig. 3a. Apparently, only the range  $0 \leq \delta \leq 3 \text{ Å}$  is compatible with the experiment and the associated error bars, indicating a screening within one atomic layer of Mg.

To scrutinize the origin of this rapid interfacial screening, we used time-dependent density functional theory (TDDFT)<sup>30</sup> to calculate  $E_L(z, t)$  for the Mg/W(110) systems. The surface-normal component of the incident laser field induces a polarization charge layer at the metal surface that shields the interior of the solid against the external electric field. The centroid of the induced screening charge density defines the exact position  $z_{\text{im}}$  of the image plane, which marks the microscopic onset of the local screening process<sup>30</sup> (see Fig. 3b). Both the positions  $z_{\text{im}, n}$  and the width  $\delta_n$  of the induced screening charge are found to be almost independent of the number  $n$  of Mg layers. The key finding is that the laser field is already fully screened at the plane defined by the centre of the atoms of the topmost layer for all Mg/W(110) systems, in agreement with the conclusion drawn from the comparison of our experimental data with the phenomenological modelling.

Finally, we incorporated the abrupt screening of the streaking field at the surface revealed by TDDFT in a full three-dimensional streaking simulation<sup>22</sup> of the electron propagation in Mg/W(110) (see Supplementary Information for a detailed description). Similar to the above-mentioned one-dimensional model, wave packet propagation is simulated by the transport of an ensemble of point-like charges taking stochastic inelastic and elastic scattering events into account. The time delays  $\Delta\tau[4f-2p]$  predicted by these calculations (red squares in Fig. 3a) are in good agreement with the experiment.

Compared to the core-level photoemission time delay  $\Delta\tau[W 4f-2p]$ , the temporal shift of the conduction band emission  $\Delta\tau[\text{CB}-2p]$  (Fig. 3c) is distinctly smaller and exhibits a strikingly different dependence on the number of Mg adlayers. A detailed analysis of  $\Delta\tau[\text{CB}-2p]$  within electron transport models is complicated by different (spectrally unresolved) contributions of W(110)- and Mg-derived states to the joint CB feature at  $E_{\text{kin}} \approx 115 \text{ eV}$ . However, by weighting the excitation probabilities from these different initial states according to their atomic photoexcitation cross-sections (see Supplementary Information), we achieved good overall agreement with the experimental results, and correctly reproduced the vanishing  $\Delta\tau[\text{CB}-2p]$  time delay for bulk Mg<sup>10</sup>. This suggests that our approximate treatment of the Bloch wave packet propagation and the dielectric screening response remains valid also for more delocalized initial electronic states, in contrast to recent predictions<sup>23,24</sup>.

We emphasize that  $\Delta\tau[\text{CB}-2p]$  for  $n = 1$  is overestimated in our transport model and lies outside the experimental error margin. A deviation from the present model appears likely for the Mg/W(110) monolayer system, since strong mixing of band states at the interface may lead to a deviation of the initial-state band structure and excitation cross-sections from their bulk characteristics<sup>24</sup>. A detailed discussion of this phenomenon is beyond the scope of the present study, but indicates the potential of attosecond photoelectron spectroscopy to probe interfacial hybridization between electronic states directly in the time domain. Therefore this technique may be applied to phenomena not describable by a

simple combination of separate electronic structure models for bulk and surface layers.

This work extends the realm of attosecond spectroscopy to the direct observation of atomic-scale propagation and damping of electronic and optical wave packets at solid surfaces. The resultant insight into attosecond temporal and—simultaneously—ångström spatial dimensions opens the door for understanding and exploring electron transport phenomena on the atomic scale and the dielectric response of solids at optical frequencies. Applied to overlayer materials with non-free-electron-like positive-energy states, such studies will shed light on whether stationary band structure can be used to predict atomic-scale electron propagation on ultrashort timescales. Extrapolation of coverage-dependent streaking spectroscopy to the sub-monolayer regime will provide access to absolute photoemission times and possible intrinsic (atomic) retardation effects in the photoemission process. Beyond addressing these fundamental questions, attosecond electron transport chronoscopy may prove instrumental in advancing electronic and photonic circuits towards atomic dimensions.

Received 15 September; accepted 14 November 2014.

- Bloch, F. Über die Quantenmechanik der Elektronen in Kristallgittern. *Z. Phys.* **52**, 555–600 (1929).
- Leo, K., Bolivar, P. H., Brüggemann, F., Schwedler, R. & Köhler, K. Observation of Bloch oscillations in a semiconductor superlattice. *Solid State Commun.* **84**, 943–946 (1992).
- Hentschel, M. et al. Attosecond metrology. *Nature* **414**, 509–513 (2001).
- Schiller, F., Heber, M., Servedio, V. D. P. & Laubschat, C. Electronic structure of Mg: from monolayers to bulk. *Phys. Rev. B* **70**, 125106 (2004).
- Gremillet, L. et al. Time-resolved observation of ultrahigh intensity laser-produced electron jets propagating through transparent solid targets. *Phys. Rev. Lett.* **83**, 5015–5018 (1999).
- Sha, W., Norris, T. B., Schaff, W. J. & Meyer, K. E. Time-resolved ballistic acceleration of electrons in a GaAs quantum-well structure. *Phys. Rev. Lett.* **67**, 2553–2556 (1991).
- Sha, W., Rhee, J.-k., Member, S., Norris, T. B. & Schaff, W. J. Transient carrier and field dynamics in quantum-well parallel transport: from the ballistic to the quasi-equilibrium regime. *IEEE J. Quantum Electron.* **28**, 2445–2455 (1992).
- Shaner, E. & Lyon, S. Picosecond time-resolved two-dimensional ballistic electron transport. *Phys. Rev. Lett.* **93**, 037402 (2004).
- Cavalieri, A. L. et al. Attosecond spectroscopy in condensed matter. *Nature* **449**, 1029–1032 (2007).
- Neppi, S. et al. Attosecond time-resolved photoemission from core and valence states of magnesium. *Phys. Rev. Lett.* **109**, 22–26 (2012).
- Schultze, M. et al. Delay in photoemission. *Science* **328**, 1658–1662 (2010).
- Aballe, L., Barinov, A., Locatelli, A., Montes, T. O. & Kiskinova, M. Initial stages of heteroepitaxial Mg growth on W(110): early condensation, anisotropic strain, and self-organized patterns. *Phys. Rev. B* **75**, 115411 (2007).
- Mahan, G. D. Theory of photoemission in simple metals. *Phys. Rev. B* **2**, 4334–4350 (1970).
- Berglund, C. N. & Spicer, W. E. Photoemission studies of copper and silver: theory. *Phys. Rev.* **136**, A1030–A1044 (1964).
- Feibelman, P. J. & Eastman, D. E. Photoemission spectroscopy—correspondence between quantum theory and experimental phenomenology. *Phys. Rev. B* **10**, 4932–4947 (1974).
- Borisov, G., Sánchez-Portal, D., Kazansky, K. & Echenique, P. M. Resonant and nonresonant processes in attosecond streaking from metals. *Phys. Rev. B* **87**, 121110 (2013).
- Krasovskii, E. E. Attosecond spectroscopy of solids: streaking phase shift due to lattice scattering. *Phys. Rev. B* **84**, 195106 (2011).
- Liao, Q. & Thumm, U. Attosecond time-resolved photoelectron dispersion and photoemission time delays. *Phys. Rev. Lett.* **112**, 023602 (2014).
- Kazansky, A. K. & Echenique, P. M. One-electron model for the electronic response of metal surfaces to subfemtosecond photoexcitation. *Phys. Rev. Lett.* **102**, 177401 (2009).
- Zhang, C. H. & Thumm, U. Attosecond photoelectron spectroscopy of metal surfaces. *Phys. Rev. Lett.* **102**, 123601 (2009).
- Zhang, C. H. & Thumm, U. Effect of wave-function localization on the time delay in photoemission from surfaces. *Phys. Rev. A* **84**, 033401 (2011).
- Lemell, C., Sölleder, B., Tökési, K. & Burgdörfer, J. Simulation of attosecond streaking of electrons emitted from a tungsten surface. *Phys. Rev. A* **79**, 62901 (2009).
- Zhang, C. H. & Thumm, U. Probing dielectric-response effects with attosecond time-resolved streaked photoelectron spectroscopy of metal surfaces. *Phys. Rev. A* **84**, 1–7 (2011).
- Vinogradov, N., Marchenko, D., Shikin, A., Adamchuk, V. & Rader, O. Size effects in ultrathin Mg/W(110) films: quantum electronic states. *Phys. Solid State* **51**, 179–188 (2009).
- Kroemer, H. On the group velocity of Bloch waves. *Proc. IEEE* **63**, 988 (1975).



26. Bartynski, R. A., Gaylord, R. H., Gustafsson, T. & Plummer, E. W. Angle-resolved photoemission study of the surface and bulk electronic structure of Mg(0001) and Mg(112-bar0). *Phys. Rev. B* **33**, 3644–3665 (1986).
27. Tanuma, S., Powell, C. J. & Penn, D. R. Calculations of electron inelastic mean free paths. IX. Data for 41 elemental solids over the 50 eV to 30 keV range. *Surf. Interf. Anal.* **43**, 689–713 (2011).
28. Lang, N. D. & Kohn, W. Theory of metal surfaces: induced surface charge and image potential. *Phys. Rev. B* **7**, 3541–3550 (1973).
29. Liebsch, A. Electronic screening at metal surfaces and the connection with physical phenomena. *Phys. Scr.* **35**, 354 (1987).
30. Wachter, G. *et al.* Electron rescattering at metal nanotips induced by ultrashort laser pulses. *Phys. Rev. B* **86**, 035402 (2012).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This research was supported by the Munich-Centre for Advanced Photonics. C.L., G.W. and J.B. acknowledge support by the FWF special research programs SFB-041 (ViCoM) and SFB-049 (NextLite) and project P21141-N16. G.W. is supported by the International Max Planck Research School for Advanced Photon Science (IMPRS-APS). R.K. acknowledges an ERC Starting Grant. Calculations have

been performed on the Vienna Scientific Cluster. S.N. and P.F. thank the Helmholtz Zentrum Berlin for support. We thank P. Echenique, E. E. Krasovskii, A. Kazansky and A. D. Sanchez-Portal for discussions.

**Author Contributions** S.N. conceived the material system for this study and performed preparatory experiments. S.N., A.L.C., P.F., E.M., R.E. and R.K. designed and developed the experiment. S.N., R.E. and A.L.C. performed the measurements (with the assistance of E.M., M.J. and E.M.B.). S.N. and R.E. analysed the data. C.L. and S.N. performed the ballistic electron simulations. G.W. and C.L. performed the TDDFT calculations. M.H. and U.K. developed and prepared the XUV multilayer optics. S.N., R.E., C.L., J.B. and R.K. wrote the manuscript with input from the other authors. R.K. and F.K. initiated the project and R.K., F.K. and P.F. supervised the project. All authors discussed the results and conclusions drawn from them.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.N. (snepp1@lbl.gov) and R.K. (reinhard.kienberger@tum.de).

# Modulation of hydrophobic interactions by proximally immobilized ions

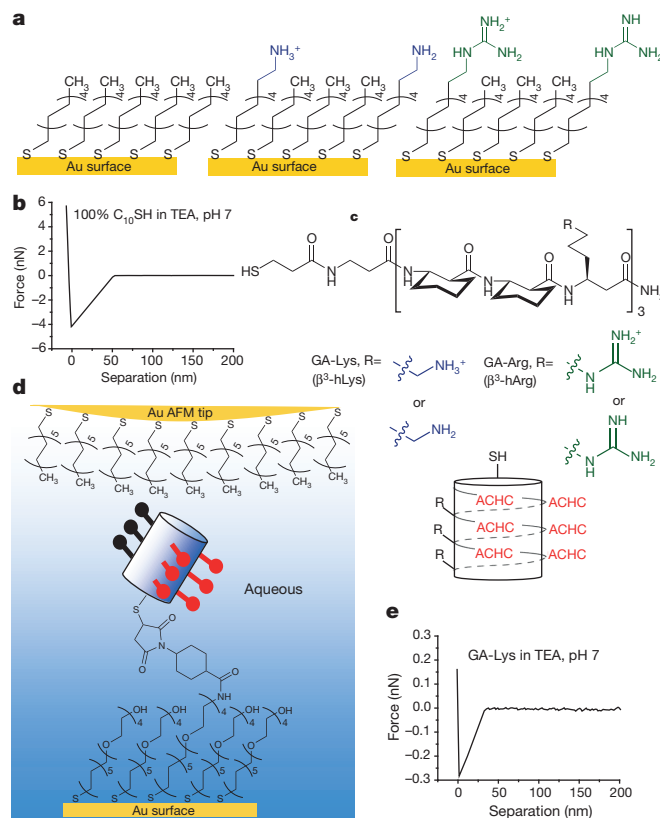
C. Derek Ma<sup>1</sup>, Chenxuan Wang<sup>1,2</sup>, Claribel Acevedo-Vélez<sup>1</sup>, Samuel H. Gellman<sup>2</sup> & Nicholas L. Abbott<sup>1</sup>

The structure of water near non-polar molecular fragments or surfaces mediates the hydrophobic interactions that underlie a broad range of interfacial, colloidal and biophysical phenomena<sup>1–4</sup>. Substantial progress over the past decade has improved our understanding of hydrophobic interactions in simple model systems<sup>1,5–10</sup>, but most biologically and technologically relevant structures contain non-polar domains in close proximity to polar and charged functional groups. Theories and simulations exploring such nanometre-scale chemical heterogeneity find it can have an important effect<sup>8,10–12</sup>, but the influence of this heterogeneity on hydrophobic interactions has not been tested experimentally. Here we report chemical force microscopy measurements on alkyl-functionalized surfaces that reveal a dramatic change in the surfaces' hydrophobic interaction strengths on co-immobilization of amine or guanidine groups. Protonation of amine groups doubles the strength of hydrophobic interactions, and guanidinium groups eliminate measurable hydrophobic interactions in all pH ranges investigated. We see these divergent effects of proximally immobilized cations also in single-molecule measurements on conformationally stable  $\beta$ -peptides with non-polar subunits located one nanometre from either amine- or guanidine-bearing subunits. Our results demonstrate the importance of nanometre-scale chemical heterogeneity, with hydrophobicity not an intrinsic property of any given non-polar domain but strongly modulated by functional groups located as far away as one nanometre. The judicious placing of charged groups near hydrophobic domains thus provides a strategy for tuning hydrophobic driving forces to optimize molecular recognition or self-assembly processes.

To quantify the hydrophobic component of interactions between two chemically defined surfaces, we compare adhesive forces measured in either aqueous triethanolamine (TEA, 10 mM) buffer or a mixture of 40 vol% aqueous TEA buffer and 60 vol% methanol. We measured the mean adhesive ('pull-off') force between the tip (gold-coated and functionalized with a monolayer of  $C_{12}H_{21}SH$ ) of an atomic force microscope (AFM) and a  $C_{10}H_{19}SH$  monolayer formed on a gold film (Figs 1a, b and 2a) to be  $4.0 \pm 0.3$  nN in the aqueous TEA buffer, independent of pH between 7 and 10.5 (Fig. 2a, b and Extended Data Figs 1, 2). Johnson–Kendall–Roberts theory<sup>13</sup> and measurements of the work of adhesion (obtained from contact angle measurements and use of Young's equation<sup>14</sup>) confirmed the magnitude of the measured force to be consistent with previous reports of hydrophobic interactions between non-polar surfaces (Methods). Addition of 60 vol% methanol to the aqueous TEA buffer reduced the measured adhesive force between the AFM tip and the  $C_{10}H_{19}SH$  monolayer from  $4.0 \pm 0.3$  to  $2.1 \pm 0.2$  nN, consistent with past studies that have shown that addition of methanol to aqueous solutions disrupts the structure of water near non-polar surfaces and diminishes or eliminates hydrophobic interactions<sup>15–19</sup>. The adhesive force measured in pure methanol is  $1.8 \pm 0.2$  nN (Fig. 2b), indicating that addition of 60 vol% methanol to aqueous TEA eliminates approximately 85% of the total hydrophobic adhesion. The same methanol proportion was reported to disrupt hydrophobically driven association of amphiphilic oligopeptides<sup>15,17</sup>. Although the force of  $2.1 \pm 0.2$  nN measured

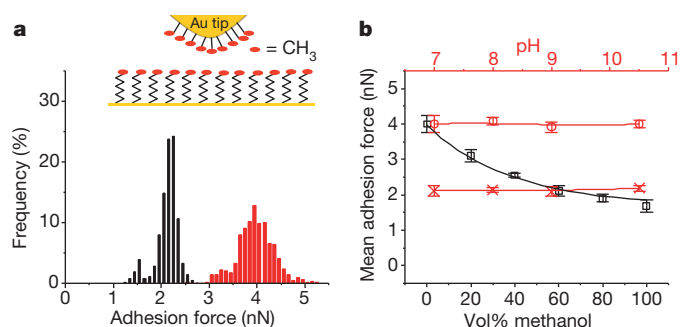
between the AFM tip and the  $C_{10}H_{19}SH$  surface in 60 vol% methanol (Fig. 2b) thus contains some residual hydrophobic interactions ( $0.3 \pm 0.2$  nN; see Methods), it arises mainly from van der Waals interactions ( $1.8 \pm 0.2$  nN).

We next used 60 vol% methanol addition to identify hydrophobic components of the adhesive interaction between an alkyl-terminated AFM tip and either  $AmC_{11}H_{22}SH$  monolayers or mixed monolayers comprised of ammonium (Am)- and alkyl-tipped thiols (40%  $AmC_{11}H_{22}SH$  and 60%  $C_{10}H_{19}SH$ ; Fig. 1a and Extended Data Fig. 3). Adhesion involving the pure  $AmC_{11}H_{22}SH$  monolayers (Fig. 3a, c and Extended



**Figure 1 | Experimental systems used to investigate the effects of immobilized charge on hydrophobic interactions.** **a**, Single-component and mixed, self-assembled monolayers presenting alkyl, Am and Gdm functional groups. **b**, Representative pull-off force curve measured when retracting an alkyl-terminated AFM tip from a  $C_{10}H_{21}SH$  monolayer in 10 mM TEA at pH 7. **c**, Linear (top) and helical (bottom) representations of the globally amphiphilic (GA)  $\beta$ -peptides used in this study. **d**, Schematic illustration of the interaction of an alkyl-terminated AFM tip with an immobilized helical  $\beta$ -peptide with side chains colour-coded to match **c**. **e**, Representative pull-off force curve for an alkyl-terminated AFM tip retracting from a surface presenting GA-Lys immobilized on a monolayer displaying 0.1% hydroxyl-terminated and amine-terminated tetraethylene glycol (Methods) in 10 mM TEA at pH 7.

<sup>1</sup>Department of Chemical and Biological Engineering, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, Wisconsin 53706, USA. <sup>2</sup>Department of Chemistry, University of Wisconsin-Madison, 1101 University Avenue, Madison, Wisconsin 53706, USA.

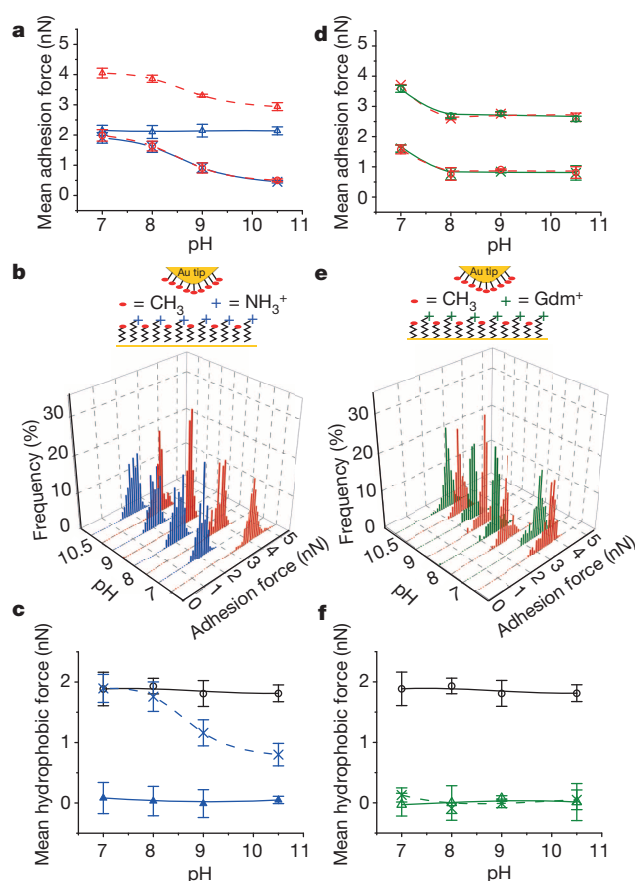


**Figure 2 | Validation of measurement of hydrophobic interaction by addition of methanol.** **a**, Adhesion force histograms for C<sub>10</sub>H<sub>21</sub>SH monolayers measured at pH 7 (red, in 10 mM TEA; black, in 60 vol% methanol). **b**, Influence of methanol added to 10 mM TEA (black squares, pH 7) and of pH (red circles, in TEA; red crosses, in 60 vol% methanol) on the mean adhesion forces calculated from histograms of C<sub>10</sub>H<sub>21</sub>SH monolayers. Adhesion force histograms were obtained using an alkyl-terminated AFM tip and 1,057–4,306 pull-off curves from 3–8 independent samples (Extended Data Tables 1 and 2). Data show mean  $\pm$  s.e.m. Lines are drawn to guide the eye.

Data Fig. 1b) was not affected by addition of 60 vol% methanol, indicating that it has no hydrophobic component. pH and salt concentration changes both altered the attractive interaction strength (Extended Data Figs 4 and 5), indicating that adhesion is affected by a surface concentration of ammonium cations that increases with decreasing pH and also by the negative surface charge on the AFM tip (Methods). In contrast, force measurements with the AmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH mixed monolayers revealed evidence of hydrophobic interactions: the pull-off force measured between the AFM tip and the mixed monolayer increased from  $2.9 \pm 0.1$  to  $4.1 \pm 0.2$  nN as the pH of the aqueous TEA buffer decreased from 10.5 to 7 (Fig. 3a, b), and stayed at  $2.1 \pm 0.4$  nN independently of solution pH in the presence of 60 vol% methanol (Fig. 3a, b). Because the decrease in adhesion force on addition of methanol identifies a hydrophobic contribution, these data reveal that protonation of the amine groups with decreasing pH doubles the strength of the hydrophobic interaction (from  $0.80 \pm 0.2$  to  $1.9 \pm 0.2$  nN) (Fig. 3c).

We next replaced Am by guanidinium (Gdm) in the monolayers (Fig. 1a), and found that addition of 60 vol% methanol to aqueous TEA revealed no measurable hydrophobic interaction between pure GdmC<sub>11</sub>H<sub>22</sub>SH monolayers and the AFM tip for pH values between 10.5 and 7 (Fig. 3d). This parallels the behaviour of the AmC<sub>11</sub>H<sub>22</sub>SH monolayers (Extended Data Fig. 1b; see also Extended Data Fig. 1c and Methods for a discussion of the orientation-dependent interactions of Gdm). But in contrast to results with mixed AmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH monolayers, the addition of methanol to TEA buffer (from pH 10.5 to 7) also had no effect on adhesive interactions between the AFM tip and the mixed GdmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH surfaces (Fig. 3d, e). Proximally immobilized Gdm ions thus seem to eliminate measurable hydrophobic interactions between the non-polar domains of the mixed monolayer and the alkyl-terminated AFM tip (Fig. 3f; see Methods for additional discussion of effects of pH).

If immobilized cationic groups can substantially alter hydrophobic interactions of neighbouring non-polar domains in an ion-specific manner, and given the prevalence of lysine and arginine residues in proteins (Am and, respectively, Gdm group in side chains), we proposed that modulation of local hydrophobic interactions by neighbouring charged groups may be important in peptidic systems. We therefore performed force measurements between an alkyl-terminated AFM tip and single  $\beta$ -peptide molecules presented from surfaces<sup>15</sup> (Fig. 1d, e). We used  $\beta$ -peptides known to adopt a helical conformation that is globally amphiphilic, with non-polar side chains of *trans*-2-aminocyclohexanecarboxylic acid (ACHC) residues and cationic side chains of either  $\beta^3$ -homolysine ( $\beta^3$ -hLys) or  $\beta^3$ -homoarginine ( $\beta^3$ -hArg) segregated onto opposite faces of the helix (Fig. 1c). This functional group segregation leads to a well-defined, ACHC-rich non-polar domain  $\sim 1$  nm<sup>2</sup> in size and a patch of



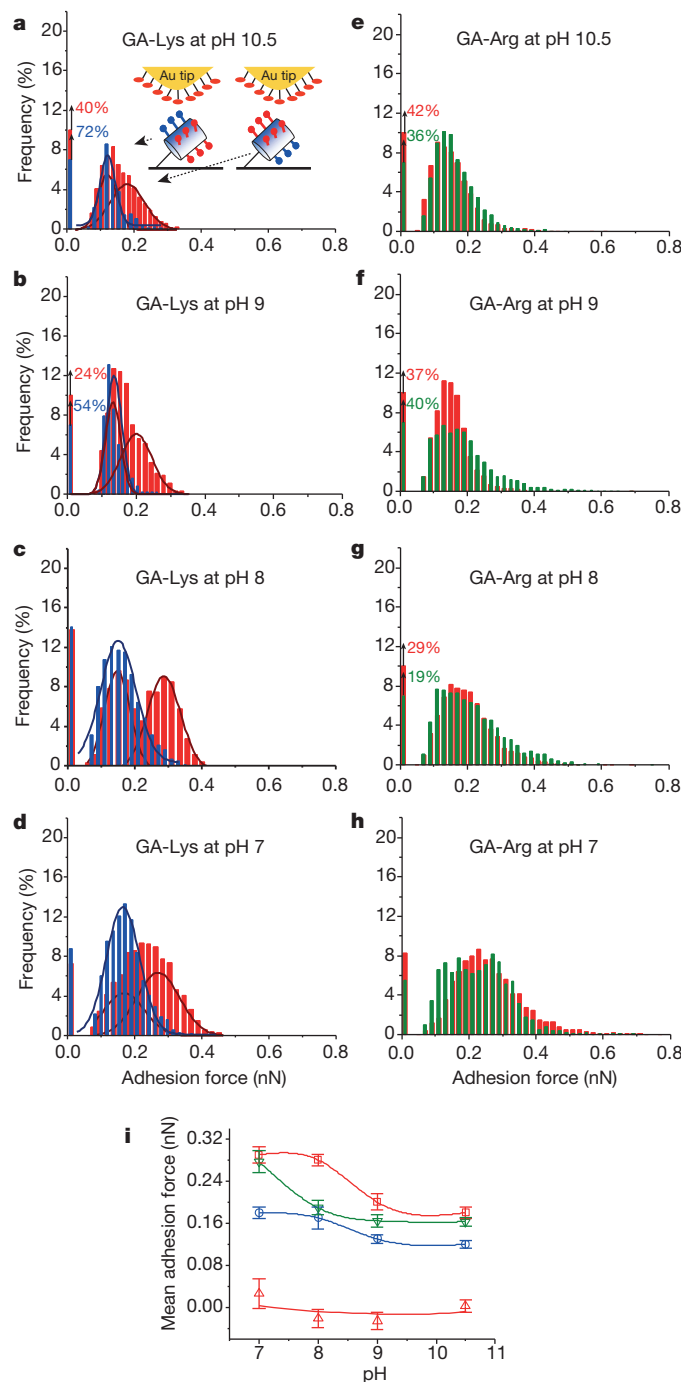
**Figure 3 | Influence of immobilized Am or Gdm on hydrophobic interactions at monolayer surfaces.** **a**, pH dependence of mean adhesion forces measured using AmC<sub>11</sub>H<sub>22</sub>SH in either 10 mM TEA (red circles) or 60 vol% methanol (blue crosses), and 40% AmC<sub>11</sub>H<sub>22</sub>SH–60% C<sub>10</sub>H<sub>21</sub>SH in either 10 mM TEA (red triangles) or 60 vol% methanol (blue triangles). **b**, Histograms of adhesion forces measured using 40% AmC<sub>11</sub>H<sub>22</sub>SH–60% C<sub>10</sub>H<sub>21</sub>SH as a function of pH (red, in TEA; blue, in 60 vol% methanol). **c**, Hydrophobic contribution to the mean adhesion forces measured using AmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH (blue crosses), AmC<sub>11</sub>H<sub>22</sub>SH (blue triangles) or C<sub>10</sub>H<sub>21</sub>SH (black circles) monolayers. **d**, Mean adhesion force measured using GdmC<sub>11</sub>H<sub>22</sub>SH in TEA (red circles) or 60 vol% methanol (green crosses), and 40% GdmC<sub>11</sub>H<sub>22</sub>SH–60% C<sub>10</sub>H<sub>21</sub>SH in TEA (red crosses) or 60 vol% methanol (blue crosses). **e**, Histograms of adhesion forces measured using 40% GdmC<sub>11</sub>H<sub>22</sub>SH–60% C<sub>10</sub>H<sub>21</sub>SH as a function of pH (red, in TEA; green, in 60 vol% methanol). **f**, Hydrophobic contribution to the mean adhesion forces measured using GdmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH (green crosses), GdmC<sub>11</sub>H<sub>22</sub>SH (green triangles) or C<sub>10</sub>H<sub>21</sub>SH (black circles). Adhesion force histograms were obtained using an alkyl-terminated AFM tip and 922–3,218 pull-off force curves from 3–6 independent samples (Extended Data Table 2). Data show mean  $\pm$  s.e.m. Lines are drawn to guide the eye.

three cationic  $\beta^3$ -residue side chains immobilized  $\sim 1$  nm from the non-polar domain. Short  $\beta$ -peptides containing  $\geq 50\%$  ACHC exhibit very stable 14-helical conformations in aqueous solution as well as in aqueous methanol<sup>17,20</sup>, thus allowing identification of hydrophobic interactions by the addition of methanol. Conventional peptides, comprised of  $\alpha$ -amino acid residues, lack sufficient  $\alpha$ -helix stability to support this experimental design<sup>21</sup>.

We immobilized  $\beta$ -peptides onto monolayers terminated in tetraethylene glycol at a low surface density to allow measurement of the interaction of single  $\beta$ -peptide molecules with an alkyl-terminated AFM tip<sup>15</sup> (Methods and Extended Data Fig. 6). Adhesive force distributions obtained for the lysine-containing  $\beta$ -peptide (GA-Lys) in aqueous buffer were broad and changed with pH (Fig. 4a–d), but narrowed in the presence of 60 vol% methanol. The latter could be fitted to a single Gaussian, whereas two Gaussian functions—one identical to the Gaussian used



to fit the 60 vol% methanol data—were needed to provide a good fit to the data measured in aqueous buffer. We attribute the two sets of interactions undergone by GA-Lys in aqueous buffer to two modes of interaction between a  $\beta$ -peptide and the hydrophobic tip (Fig. 4a, inset cartoon),



**Figure 4 | Influence of lysine and arginine side chains on hydrophobic interactions involving oligopeptides.** **a–h**, Histograms of adhesion forces measured between alkyl-terminated AFM tips and immobilized  $\beta$ -peptides. The surfaces presented either GA-Lys (**a–d**) or GA-Arg (**e–h**), and the forces are reported as functions of pH in either 10 mM TEA (red) or 60 vol% methanol (blue, GA-Lys; green, GA-Arg). The inset in **a** shows schematic illustrations of the interaction between the AFM tip and GA-Lys. **i**, Dependence on pH of the hydrophobic (red squares, GA-Lys; red triangles, GA-Arg) and non-hydrophobic (blue circles, GA-Lys; green triangles, GA-Arg) components of the interaction of  $\beta$ -peptides with the AFM tip. Adhesion force histograms were obtained using 1,542–5,004 pull-off force curves from 4–7 independent samples (Extended Data Table 2). Data show mean  $\pm$  s.e.m. Lines are drawn to guide the eye.

with the deconvolution and force assignment supported by our past study<sup>15</sup> and by control measurements showing that 60 vol% methanol does not change the direct interactions of Am groups with the AFM tip (Fig. 2a and Extended Data Fig. 1b). Because only the weaker adhesion mode persists in the presence of 60 vol% methanol, we attribute the stronger set of adhesive events in aqueous buffer to a hydrophobic interaction arising when the  $\beta$ -peptide is oriented such that it presents a non-polar surface (ACHC residues) to the tip. The weaker adhesion events are attributed to a  $\beta$ -peptide orientation that presents largely Am surface elements and leads to interactions that are not hydrophobic in nature. This interpretation predicts that charging of the lysine side-chain amino groups should strengthen hydrophobic interactions between the non-polar domain of GA-Lys and the AFM tip, as seen in our data (Fig. 4i). For a single GA-Lys molecule, the hydrophobic component of the measured adhesion force ranges from  $0.18 \pm 0.02$  to  $0.29 \pm 0.02$  nN (Fig. 4i) and arises from a non-polar domain area of  $\sim 1$  nm<sup>2</sup>, consistent with previous single-molecule studies<sup>15,22</sup> and our monolayer measurements (Fig. 2) of adhesion forces in the range  $0.80 \pm 0.2$  to  $1.9 \pm 0.2$  nN for contact areas of 10.7 nm<sup>2</sup> (Methods).

To further explore the role of the cationic residues in modulating hydrophobic adhesion involving the non-polar domains of the  $\beta$ -peptide helix, we replaced the  $\beta^3$ -hLys residues with  $\beta^3$ -hArg residues to generate GA-Arg (Fig. 1c). In striking contrast to the behaviour of GA-Lys (Fig. 4a–d), adhesive interactions measured between GA-Arg and the AFM tip were largely unaffected by the addition of methanol (Fig. 4e–h). Although we do not interpret these measurements in the framework of multiple interaction modes between the  $\beta$ -peptide and the AFM tip<sup>23</sup> (Methods and Extended Data Fig. 7), they indicate that the adhesive interactions measured with GA-Arg in aqueous buffer are not of hydrophobic origin.

Our results, which are obtained using two independent systems in which specific cationic groups can be spatially juxtaposed with non-polar domains, the mixed monolayers and designed  $\beta$ -peptides, lead to a consistent set of conclusions regarding the effects of proximally immobilized charge on hydrophobic interactions. Specifically, both studies demonstrate that charging of immobilized Am groups strengthens hydrophobic interactions involving proximal non-polar surfaces, whereas immobilized Gdm groups diminish hydrophobic interactions to the extent that they are not measurable in our experiments. We note, however, that the spatial presentation of the non-polar and cationic groups to the AFM tip differs between the two systems, as is evident in the widths of the force distributions obtained using Am surfaces and GA-Lys (Methods and Extended Data Fig. 8).

Past studies of the effects of soluble salts on solvation of molecular surfaces have been rationalized in terms of the surface accumulation or depletion of ions<sup>24</sup>, but the effects we have documented here differ: we have evaluated the impact of cationic groups covalently immobilized very close to non-polar moieties, a situation that often occurs in proteins and other complex biomolecules yet has not previously been analysed systematically. We note that the distance between the cations and the non-polar domain in the globally amphiphilic  $\beta$ -peptides used in our study is fixed at  $\sim 1$  nm, and that the impact of cationic groups on hydrophobic interactions thus extends over at least this distance. This conclusion is generally consistent with atomistic simulations and theories predicting that non-polar surfaces can influence the structure and compressibility of water over distances of  $\sim 1$  nm (refs 8, 9), and with predictions that solvation shells are susceptible to the influence of neighbouring atoms, especially strongly charged ones<sup>12</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 1 August; accepted 27 October 2014.**

- Chandler, D. Interfaces and the driving force of hydrophobic assembly. *Nature* **437**, 640–647 (2005).

2. Meyer, E. E., Rosenberg, K. J. & Israelachvili, J. Recent progress in understanding hydrophobic interactions. *Proc. Natl Acad. Sci. USA* **103**, 15739–15746 (2006).
3. Whitesides, G. M. & Grzybowski, B. Self-assembly at all scales. *Science* **295**, 2418–2421 (2002).
4. Dyson, H. J., Wright, P. E. & Scheraga, H. A. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proc. Natl Acad. Sci. USA* **103**, 13057–13061 (2006).
5. Davis, J. G., Gierszal, K. P., Wang, P. & Ben-Amotz, D. Water structural transformation at molecular hydrophobic interfaces. *Nature* **491**, 582–585 (2012).
6. Huang, D. M. & Chandler, D. Temperature and length scale dependence of hydrophobic effects and their possible implications for protein folding. *Proc. Natl Acad. Sci. USA* **97**, 8324–8327 (2000).
7. Li, I. T. S. & Walker, G. C. Signature of hydrophobic hydration in a single polymer. *Proc. Natl Acad. Sci. USA* **108**, 16527–16532 (2011).
8. Acharya, H., Vembanur, S., Jamadagni, S. N. & Garde, S. Mapping hydrophobicity at the nanoscale: applications to heterogeneous surfaces and proteins. *Faraday Discuss.* **146**, 353–365 (2010).
9. Patel, A. J., Varilly, P. & Chandler, D. Fluctuations of water near extended hydrophobic and hydrophilic surfaces. *J. Phys. Chem. B* **114**, 1632–1637 (2010).
10. Patel, A. J. *et al.* Sitting at the edge: how biomolecules use hydrophobicity to tune their interactions and function. *J. Phys. Chem. B* **116**, 2498–2503 (2012).
11. Giovambattista, N., Debenedetti, P. G. & Rossky, P. J. Hydration behavior under confinement by nanoscale surfaces with patterned hydrophobicity and hydrophilicity. *J. Phys. Chem. C* **111**, 1323–1332 (2007).
12. Li, L., Fennell, C. J. & Dill, K. A. Field-SEA: a model for computing the solvation free energies of nonpolar, polar, and charged solutes in water. *J. Phys. Chem. B* **118**, 6431–6437 (2014).
13. Johnson, K. L., Kendall, K. & Roberts, A. D. Surface energy and contact of elastic solids. *Proc. R. Soc. Lond. A* **324**, 301–313 (1971).
14. Young, T. An essay on the cohesion of fluids. *Philos. Trans. R. Soc. Lond.* **95**, 65–87 (1805).
15. Acevedo-Vélez, C., Andre, G., Dufrene, Y. F., Gellman, S. H. & Abbott, N. L. Single-molecule force spectroscopy of beta-peptides that display well-defined three-dimensional chemical patterns. *J. Am. Chem. Soc.* **133**, 3981–3988 (2011).
16. Hwang, S., Shao, Q., Williams, H., Hilty, C. & Gao, Y. Q. Methanol strengthens hydrogen bonds and weakens hydrophobic interactions in proteins - a combined molecular dynamics and NMR study. *J. Phys. Chem. B* **115**, 6653–6660 (2011).
17. Pomerantz, W. C., Grygiel, T. L. R., Lai, J. R. & Gellman, S. H. Distinctive circular dichroism signature for 14-helix-bundle formation by beta-peptides. *Org. Lett.* **10**, 1799–1802 (2008).
18. Vezenov, D. V., Zhuk, A. V., Whitesides, G. M. & Lieber, C. M. Chemical force spectroscopy in heterogeneous systems: intermolecular interactions involving epoxy polymer, mixed monolayers, and polar solvents. *J. Am. Chem. Soc.* **124**, 10578–10588 (2002).
19. Wang, J. L., Li, Z. L., Yoon, R. H. & Eriksson, J. C. Surface forces in thin liquid films of *n*-alcohols and of water-ethanol mixtures confined between hydrophobic surfaces. *J. Colloid Interface Sci.* **379**, 114–120 (2012).
20. Raguse, T. L., Lai, J. R. & Gellman, S. H. Environment-independent 14-helix formation in short  $\beta$ -peptides: striking a balance between shape control and functional diversity. *J. Am. Chem. Soc.* **125**, 5592–5593 (2003).
21. Chakrabarty, A. & Baldwin, R. L. Stability of  $\alpha$ -helices. *Adv. Protein Chem.* **46**, 141–176 (1995).
22. Hinterdorfer, P. & Dufrene, Y. F. Detection and localization of single molecular recognition events using atomic force microscopy. *Nature Methods* **3**, 347–355 (2006).
23. Godawat, R., Jamadagni, S. N. & Garde, S. Unfolding of hydrophobic polymers in guanidinium chloride solutions. *J. Phys. Chem. B* **114**, 2246–2254 (2010).
24. Lo Nostro, P. & Ninham, B. W. Hofmeister phenomena: an update on ion specificity in biology. *Chem. Rev.* **112**, 2286–2322 (2012).

**Acknowledgements** This research was supported by the Wisconsin Nanoscale Science and Engineering Center (NSF grant DMR-0832760). Use of facilities supported by the Wisconsin Materials Research Science and Engineering Center is also acknowledged (NSF grant DMR-1121288).

**Author Contributions** C.D.M. and C.A.-V. synthesized, characterized and performed all force measurements involving oligopeptides. C.W. prepared samples and performed all measurements involving monolayers. S.H.G. and N.L.A. were involved in study design and data interpretation, and wrote the manuscript. All authors discussed the results and commented on the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.H.G. ([gellman@chem.wisc.edu](mailto:gellman@chem.wisc.edu)) or N.L.A. ([abbott@engr.wisc.edu](mailto:abbott@engr.wisc.edu)).

## METHODS

**Materials.** Tetraethylene glycol thiols terminated in hydroxyl (EG4) or amine groups (EG4N) along with 11-guanidinoundecanethiol (GdmC<sub>11</sub>SH) were purchased from Prochimia. 1-Decanethiol (C<sub>10</sub>SH, 96%), 1-dodecanethiol (C<sub>12</sub>SH, 98%), 1-hexadecanethiol (C<sub>16</sub>SH, 95%), triethanolamine HCl (TEA, 99%) and methanol (anhydrous, 99.8%) were purchased from Sigma-Aldrich. Sulphosuccinimidyl-4-(*N*-maleimidomethyl)cyclohexane-1-carboxylate (SSMCC) was purchased from Pierce Biotechnology. 11-Aminoundecanethiol (AmC<sub>11</sub>SH) was purchased from Dojindo Molecular Technologies. Ethanol (reagent, anhydrous, denatured) used for preparation of thiol solutions and sodium chloride (99.0%) were purchased from Sigma-Aldrich. Ethanol (anhydrous, 200 proof) used for rinsing was purchased from Decon Labs. De-ionized water used in this study had a resistivity of 18.2 MΩ cm. All chemicals were used as received and without any further purification. The AFM tips used in this study (triangular in shape, nominal spring constant of 0.01 N m<sup>-1</sup> or 0.03 N m<sup>-1</sup>) were purchased from Bruker. Silicon wafers were purchased from Silicon Sense.

**Preparation of β-peptide-decorated surfaces.** β-Peptide oligomers were synthesized by solid-phase methods as described elsewhere<sup>25</sup> and immobilized as detailed previously<sup>15</sup>. In brief, we immobilized the β-peptides on mixed monolayers terminated in EG4 and EG4N groups, using a mole fraction of EG4N of 0.001 to achieve a low surface density of the β-peptides. This approach enabled us to measure adhesive interactions between single β-peptide molecules and the AFM tip; prior work has shown that EG4N-EG4 mixed monolayers do not generate measurable adhesive forces with hydrophobic AFM tips in aqueous TEA<sup>15</sup>.

**Preparation of pure or mixed-component monolayer surfaces.** Monolayers of organothiols were prepared by immersing small pieces of gold-coated silicon wafers into ethanolic solutions containing 1 mM of the appropriate thiol components and incubating for 18 h. On removal from solution, hydrophilic substrates were rinsed thoroughly with ethanol and water, dried with nitrogen and stored in TEA buffer (10 mM, pH 7) until AFM adhesion measurements were performed. Hydrophobic substrates were rinsed thoroughly with ethanol, dried with nitrogen and used immediately for AFM measurement. The mixed monolayers shown in Fig. 1a were prepared by co-adsorption of ω-functionalized undecanethiols (AmC<sub>11</sub>SH or GdmC<sub>11</sub>SH at 0.4 mM) and decanethiol (C<sub>10</sub>SH at 0.6 mM).

**Preparation of chemically functionalized AFM tips.** Triangular-shaped cantilevers with nominal spring constants of 0.01 N m<sup>-1</sup> were used for experiments involving β-peptides, whereas cantilevers with nominal spring constants of 0.03 N m<sup>-1</sup> were used for experiments involving monolayers unless otherwise specified. The spring constants of the cantilevers were calibrated using Sader's method on a PCM-90 Spring Constant Calibration Module (Novascan Technologies) and determined to be respectively ~0.01 N m<sup>-1</sup> and ~0.03 N m<sup>-1</sup>. AFM tips were coated with a 2 nm layer of titanium and a 20 nm layer of gold by physical deposition using an electron beam evaporator. Following the coating by gold, the tips were immersed in a 1 mM ethanolic solution of C<sub>12</sub>SH and incubated overnight. On removal from solution, the tips were rinsed with ethanol, dried with a gentle stream of nitrogen and immediately transferred to the AFM fluid cell.

**Characterization of mixed monolayers.** Ellipsometry was used to confirm formation of monolayers of organothiols on gold films (Extended Data Fig. 3a), and X-ray photoelectron spectroscopy was used to determine the compositions of the mixed monolayers shown in Extended Data Fig. 3b. The results shown in Extended Data Fig. 3 reveal that the compositions of the mixed monolayers were similar to the compositions of the thiol solutions from which the mixed monolayers were formed (C<sub>10</sub>SH mole fraction of 0.6).

**AFM force measurements.** Adhesion force measurements were performed using a Nanoscope IIIa Multimode AFM equipped with a fluid cell (Veeco Metrology Group). Triangular-shaped silicon nitride cantilevers were used and functionalized as described above. Force measurements were performed at room temperature. Force curves were recorded using a constant contact time of 500 ms and retraction and approach speeds of 1,000 nm s<sup>-1</sup>. All measurements were performed in 10 mM aqueous TEA solutions unless otherwise specified.

**Interpretation of adhesion measured between alkyl-terminated surfaces.** We measured advancing contact angles of water (θ<sub>ad</sub>) on monolayers formed from C<sub>10</sub>H<sub>21</sub>SH and C<sub>12</sub>H<sub>23</sub>SH to be 108 ± 3° and 107 ± 2°, respectively, consistent with formation of hydrophobic surfaces. We also confirmed that the addition of TEA to water does not measurably influence the water-mediated adhesion between the two hydrophobic surfaces (Extended Data Fig. 2). We used the adhesion measurements shown in Extended Data Fig. 2 in combination with knowledge of the surface tension of water (γ<sub>liquid-vapour</sub> = 72.4 mJ m<sup>-2</sup> (ref. 18)) to estimate (i) the work of adhesion (W<sub>ad</sub>) between the alkyl-terminated AFM tip and an alkyl-terminated monolayer in water, and (ii) the apparent radius of curvature of the AFM tip in these measurements. W<sub>ad</sub> in water is related to the interfacial free energies (γ) as<sup>26</sup>

$$W_{ad} = \gamma_{\text{sample-water}} + \gamma_{\text{tip-water}} - \gamma_{\text{sample-tip}}$$

For identical surface functional groups (alkyl-terminated surfaces) on both the sample and the AFM tip, γ<sub>sample-water</sub> ≈ γ<sub>tip-water</sub> and γ<sub>sample-tip</sub> ≈ 0, and W<sub>ad</sub> thus becomes the work of cohesion<sup>26</sup> (W<sub>co</sub>):

$$W_{ad} \approx W_{co} \approx 2\gamma_{\text{sample-water}} \approx 2\gamma_{\text{tip-water}} \quad (1)$$

The interfacial energies between the surface and water in equation (1) can be obtained from contact angle measurements and use of Young's equation<sup>14</sup>:

$$\gamma_{\text{surface-water}} = \gamma_{\text{surface-vapour}} - \gamma_{\text{water-vapour}} \cos(\theta_{ad}) \quad (2)$$

By combining equations (1) and (2), and using γ<sub>surface-vapour</sub> ≈ 20 mJ m<sup>-2</sup> (ref. 27), we estimate W<sub>ad</sub> between the alkyl-terminated surface and AFM tip to be 85 mJ m<sup>-2</sup>. This value is consistent with previously reported values of W<sub>ad</sub> for hydrophobic surfaces, which range from 85 to 100 mJ m<sup>-2</sup> (ref. 28). From knowledge of W<sub>ad</sub> and the adhesion force (F<sub>ad</sub>), the Johnson-Kendall-Roberts model can be used to determine the effective AFM tip radius<sup>13</sup> (R<sub>JKR</sub>):

$$F_{ad} = \frac{3}{2} \pi R_{JKR} W_{ad} \quad (3)$$

From equation (3) (with F<sub>ad</sub> = 4.0 nN and W<sub>ad</sub> = 85 mJ m<sup>-2</sup>; see above), we estimate R<sub>JKR</sub> to be 10.7 nm, an effective radius similar to that inferred from past measurements of F<sub>ad</sub> using gold-coated AFM tips covered with monolayers of alkanethiols<sup>29,30</sup>. We note that this effective radius is smaller than the AFM tip radius determined with scanning electron microscopy (R<sub>SEM</sub> = 53 ± 5 nm). As discussed in detail elsewhere, this difference arises from the presence of gold asperities, resulting in adhesive interactions between functionalized gold domains that are smaller than R<sub>SEM</sub> (refs 30, 31).

In the main text, we report that the addition of 60 vol% methanol to the aqueous solution results in an adhesive force between alkyl-terminated surfaces of 2.1 ± 0.2 nN (Fig. 2b). In solutions containing 60 vol% methanol, we measured the AFM tip to jump into contact with the alkyl-terminated surface at a distance (D<sub>jc</sub>) of 17 ± 2 nm. We estimate the Hamaker constant (A) corresponding to the jump-to-contact distance as<sup>18</sup>

$$A = \frac{kD_{jc}^3}{3R_{SEM}} \quad (4)$$

where k = 0.01 N m<sup>-1</sup> is the AFM cantilever spring constant. From equation (4), we calculate A to be 30 × 10<sup>-20</sup> J, a value consistent with past reports of Hamaker constants for van der Waals interactions between gold surfaces (10 × 10<sup>-20</sup> to 40 × 10<sup>-20</sup> J; ref. 32). Past studies have also reported that interactions involving self-assembled monolayers formed on gold films are influenced by van der Waals forces from the underlying gold films<sup>18,33,34</sup>. Overall, our measurements of the interactions between methyl-terminated surfaces in 60 vol% methanol are consistent with a dominating influence of van der Waals forces. We note that our experimental measurements in 60 vol% methanol suggest that the adhesive force of 2.1 ± 0.2 nN comprises principally van der Waals force (1.8 ± 0.2 nN) with a small (0.3 ± 0.2 nN) residual hydrophobic contribution (Fig. 2b).

**Adhesive interactions involving ammonium-terminated surfaces (AmC<sub>11</sub>H<sub>22</sub>SH).** Figure 3a and Extended Data Fig. 1b show that adhesive interactions between AmC<sub>11</sub>H<sub>22</sub>SH monolayers and alkyl-terminated AFM tips increase in strength with decreasing pH. Here we discuss the origins of this pH dependence of the adhesive interaction. In this context, we note that past experimental studies have documented that hydrophobic surfaces acquire an excess of negative surface charge density in aqueous solution<sup>35-37</sup>. Consistent with these past studies, our electrophoretic mobility measurements of C<sub>12</sub>H<sub>23</sub>SH-functionalized gold nanoparticles (diameter of 150 nm) confirmed the presence of negative ζ-potentials in aqueous TEA (−26 ± 9 mV at pH 7 and −38 ± 9 mV at pH 9). Our adhesion force measurements are consistent with the excess negative surface charge of the hydrophobic AFM tip mediating a screened electrostatic attraction to the positively charged AmC<sub>11</sub>H<sub>22</sub>SH monolayers<sup>27,28</sup>. Specifically, we measured the presence of a long-range attraction between the approaching AFM tip and the AmC<sub>11</sub>SH surface that caused the AFM tip to jump into contact with the AmC<sub>11</sub>SH surface. The distance at which the tip jumped depended on the pH and ionic strength (Extended Data Fig. 4). An increase in ionic strength was observed to reduce the jump-to-contact distance (from 7 nm in TEA at pH 7 to 3 nm in TEA at pH 7 with 100 mM NaCl), consistent with an increase in the screening of the interaction with increasing ionic strength. We evaluated the Debye length (λ<sub>D</sub>) as<sup>26</sup>

$$\lambda_D = \sqrt{\frac{\epsilon_r \epsilon_0 k_B T}{e^2 \sum_i z_i^2 M_i}} \quad (5)$$

where ε<sub>r</sub> is the dielectric constant, ε<sub>0</sub> is the permittivity of free space, k<sub>B</sub> is the Boltzmann constant, T is the temperature, e is the electron charge, z<sub>i</sub> is the charge of the *i*th



ion and  $M_i$  is the molar concentration of the  $i$ th ion. From equation (5), we estimate  $\lambda_D$  to be 4.31 nm in TEA and 0.94 nm in TEA with 100 mM NaCl. These estimates of  $\lambda_D$  are generally consistent with the distances at which the jump to contact was measured, thus providing support for our conclusion that the long-range attraction is a consequence of the interaction of two electrical double layers. Additionally, in the absence of added salt, we observed the magnitude of the long-range attractive force to decrease with increasing pH, consistent with the effects of a decrease in the density of charge on the amine-terminated surface (Extended Data Fig. 4).

In Fig. 3c, we show that the mean hydrophobic force measured between the AFM tip and the 40% AmC<sub>11</sub>H<sub>22</sub>SH–60% C<sub>10</sub>H<sub>21</sub>SH monolayer at pH 7 is similar to the mean hydrophobic force measured between the AFM tip and the monolayer of C<sub>10</sub>H<sub>21</sub>SH. For other mixed-monolayer compositions, however, we found the adhesion force generated by the mixed monolayers at pH 7 to differ from the adhesion forces measured using C<sub>10</sub>H<sub>21</sub>SH. For example, as shown in Extended Data Fig. 5, a monolayer of 90% AmC<sub>11</sub>H<sub>22</sub>SH–10% C<sub>10</sub>H<sub>21</sub>SH at pH 7 does not generate the same magnitude of hydrophobic adhesion force as the C<sub>10</sub>H<sub>21</sub>SH monolayer.

**Adhesive interactions involving Gdm-terminated surfaces.** In Fig. 3d and Extended Data Fig. 1c, we report the pH dependence of the adhesive interaction between GdmC<sub>11</sub>H<sub>22</sub>SH surfaces and the hydrophobic AFM tip. Whereas the strength of the adhesive interaction did not significantly change between pH 10.5 and 8, on decreasing the pH from 8 to 7 a measurable increase in adhesive strength was recorded. We note that the interaction of the GdmC<sub>11</sub>H<sub>22</sub>SH surface with the hydrophobic AFM tip is more complex than the interaction of the AmC<sub>11</sub>H<sub>22</sub>SH surface with the tip because, in addition to interacting with electrical double layers, Gdm binds to non-polar domains through orientation-dependent (see below) van der Waals interactions<sup>23</sup>. The pH-dependent trends seen in Fig. 3d are also evident in the measurements shown in Fig. 4i using GA-Arg.

Inspection of Fig. 3d, f shows that GdmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH monolayers interact with the hydrophobic AFM tip more strongly than do GdmC<sub>11</sub>H<sub>22</sub>SH monolayers. Past studies have concluded that Gdm preferentially interacts with non-polar domains in a planar orientation<sup>23</sup>. We speculate that the difference in the strength of the adhesion between alkyl surfaces and either GdmC<sub>11</sub>H<sub>22</sub>SH or GdmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH surfaces stems from differences in the orientation of the Gdm groups presented by the two monolayers. For example, it is possible that within the GdmC<sub>11</sub>H<sub>22</sub>SH monolayers, the Gdm functional groups are densely packed such that they can interact with the hydrophobic surface only in a side-on orientation. In contrast, within the GdmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH surfaces, the Gdm functional groups protrude above the surface, resulting in a greater number of orientational degrees of freedom, which would allow face-on interactions that would generate larger adhesion forces.

**Interactions of single  $\beta$ -peptides with the AFM tip.** The density of surface-immobilized  $\beta$ -peptide molecules used in our study was selected by performing screening measurements at different immobilization densities. The density of surface-immobilized  $\beta$ -peptide molecules was controlled by using mixed monolayers terminated in EG4 and EG4N groups (the mole fraction of EG4N in the mixed monolayer was varied to change the immobilization density of the  $\beta$ -peptides). For ease of interpretation of our measurements, guided by our past results<sup>15</sup>, isoGA-Lys (Extended Data Fig. 6) was used in these screening measurements. We measured the mean adhesion force to increase with mole fraction of EG4N in the mixed monolayer used to immobilize the  $\beta$ -peptides when the mole fraction of EG4N was greater than 0.002 ( $0.14 \pm 0.015$  nN at a mole fraction of 0.002,  $0.7 \pm 0.05$  nN at 0.01 and  $1.1 \pm 0.08$  nN at 0.1), indicating that at immobilization densities above 0.002 the adhesive force reflects the simultaneous interaction of multiple  $\beta$ -peptides with the AFM tip. However, when the mole fraction of EG4N in the mixed monolayer used to immobilize the  $\beta$ -peptides was below 0.002, the adhesion force ceased to change with decreasing immobilization density ( $0.14 \pm 0.015$  nN at 0.002,  $0.13 \pm 0.013$  nN at 0.001), consistent with an adhesive pull-off force that reflects the interaction of a single  $\beta$ -peptide with the AFM tip. In the measurements reported in this paper, we used an immobilization density of  $\beta$ -peptides of 0.001 to obtain measurements of the interactions of single  $\beta$ -peptide molecules with the AFM tip.

To enable a comparison of the magnitude of the hydrophobic adhesion forces measured using single  $\beta$ -peptides with the magnitude of the hydrophobic adhesion forces measured using the alkyl-terminated monolayers, we used Johnson–Kendall–Roberts theory to estimate the contact radius ( $a$ ) at pull-off for the measurements using the monolayers<sup>13,38</sup>:

$$a^3 = \frac{3\pi R_{\text{JKR}}^2 W_{\text{ad}}}{2K} \quad (6)$$

where  $K$  is the elastic modulus of the contacting surfaces ( $K \approx 6.4$  GPa; refs 39, 40). By using equation (6) and the values of  $R_{\text{JKR}}$  and  $W_{\text{ad}}$  reported above, we estimate the contact area at pull-off to be  $10.7 \text{ nm}^2$ . In contrast, the size of the non-polar domain of the  $\beta$ -peptide is  $\sim 1 \text{ nm}^2$  (refs 41, 42), which is approximately ten times smaller than the contact area involved in the monolayer adhesion measurements. Consistent with this conclusion, we measured the magnitudes of the hydrophobic forces in the

monolayer experiments to lie between 0.80 and 1.9 nN (Fig. 3c), whereas in the  $\beta$ -peptide experiments they ranged from 0.16 to 0.28 nN (Fig. 4i).

**Comparing the widths of the adhesion force histograms.** As noted in the main text, differences in the spatial presentation of the non-polar and cationic groups between the  $\beta$ -peptide and the monolayer systems are evident in the force histograms obtained using GA-Lys or mixed monolayers of AmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH. Specifically, immobilized GA-Lys molecules possess rotational degrees of freedom that permit the non-polar AFM tip to interact with two distinct helical faces, that is, surfaces consisting entirely of non-polar ACHC residues or containing largely cationic residues (Fig. 4a, cartoon). This leads to two modes of interaction when measurements are performed in aqueous buffer (as described in the main text). However, when the measurements are performed in the presence of methanol, there is only one dominant mode of interaction with the AFM tip, which is reflected in a narrowing of the force histograms (from a coefficient of variation of 0.32 in aqueous buffer to 0.22 in the presence of methanol; see Extended Data Fig. 8). In contrast, the contact area between the AFM tip and a mixed monolayer of AmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH contains a statistical mixture of approximately 50 charged and non-polar groups in total, and there is thus little change in the coefficient of variation of the force histogram on addition of methanol (see Extended Data Fig. 8 and below). In contrast to observations with GA-Lys, addition of methanol during measurements involving GA-Arg did not lead to a change in the apparent number of modes of interaction of GA-Arg with the AFM tip, and the coefficients of variation of the histograms obtained with GA-Arg are thus similar with and without methanol (Extended Data Fig. 8).

To provide insight into the above-described coefficients of variation obtained from our experiments, we modelled the statistical variation of the composition of the mixed monolayers within the area of contact made with the AFM tip by using a binomial distribution. The mean ( $\mu$ ) and variance ( $\sigma^2$ ) of the number of either cationic or alkyl groups within the contact area with the AFM tip can be calculated from the total number of functional groups ( $n$ ) and the probability ( $p$ ) of finding either a cationic or alkyl functional group as

$$\mu = np$$

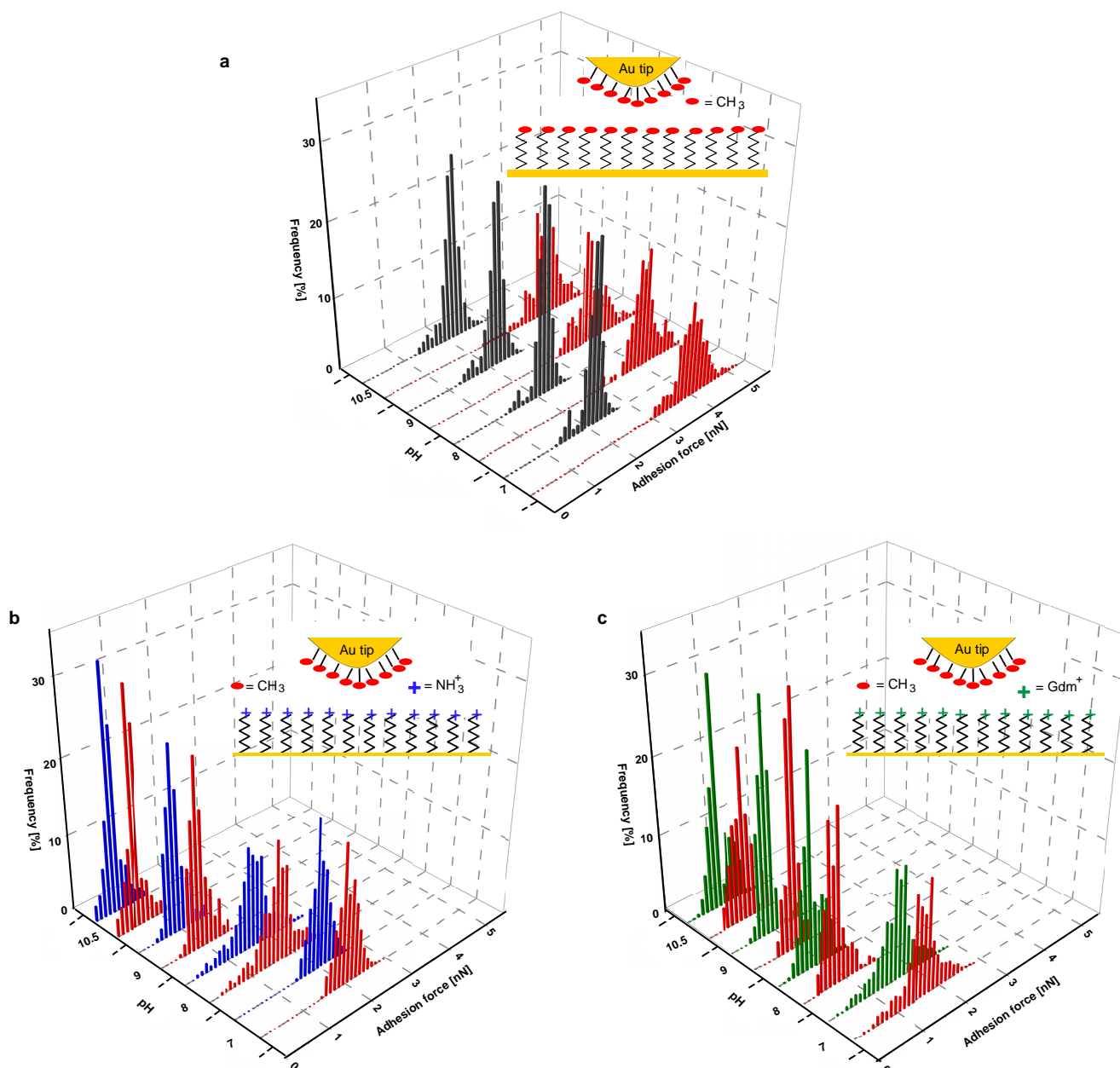
$$\sigma^2 = np(1-p)$$

Within the contact area between the AFM tip and a monolayer, we estimate the total number of groups to be  $n \approx 50$  by assuming a thiol packing density of  $0.214 \text{ nm}^2$  per thiol<sup>43</sup> and a contact area at pull-off of  $10.7 \text{ nm}^2$  (see above). We calculate the coefficient of variation for the methyl group within the mixed monolayer to be 0.12 (for  $n = 50$ ,  $P = 0.6$ ) and the coefficient of variation for the cationic group within the mixed monolayer to be 0.17 (for  $n = 50$ ,  $P = 0.4$ ). The coefficients of variation measured for the mixed monolayers (0.08 to 0.11) are comparable to the coefficients of variation calculated for a binomial distribution of cationic and non-polar units, consistent with a molecularly mixed monolayer of cationic and alkyl groups.

**Influence of the buffer anion on hydrophobic adhesion.** The measurements reported in Figs 1–4 were obtained using 10 mM TEA HCl as the buffer. We confirmed that a change in buffer did not influence the measurements of hydrophobic forces reported in this paper. Specifically, as detailed in Extended Data Fig. 7, we determined that measurements performed using 10 mM MOPS as the buffer were indistinguishable from measurements performed using TEA HCl.

- Pomerantz, W. C., Cadwell, K. D., Hsu, Y. J., Gellman, S. H. & Abbott, N. L. Sequence dependent behavior of amphiphilic  $\beta$ -peptides on gold surfaces. *Chem. Mater.* **19**, 4436–4441 (2007).
- Hiemenz, P. C. & Rajagopalan, R. *Principles of Colloid and Surface Chemistry* 3rd edn (CRC, 1997).
- Vezenov, D. V., Noy, A. & Ashby, P. Chemical force microscopy: probing chemical origin of interfacial forces and adhesion. *J. Adhes. Sci. Technol.* **19**, 313–364 (2005).
- Drellich, J., Tormoen, G. W. & Beach, E. R. Determination of solid surface tension from particle-substrate pull-off forces measured with the atomic force microscope. *J. Colloid Interface Sci.* **280**, 484–497 (2004).
- Alsteens, D., Daguerre, E., Rouxhet, P. G., Baulard, A. R. & Dufrene, Y. F. Direct measurement of hydrophobic forces on cell surfaces using AFM. *Langmuir* **23**, 11977–11979 (2007).
- Skulason, H. & Frisbie, C. D. Rupture of hydrophobic microcontacts in water: correlation of pull-off force with AFM tip radius. *Langmuir* **16**, 6294–6297 (2000).
- Awada, H., Castelein, G. & Brogly, M. Quantitative determination of surface energy using atomic force microscopy: the case of hydrophobic/hydrophobic contact and hydrophilic/hydrophilic contact. *Surf. Interface Anal.* **37**, 755–764 (2005).
- Israelachvili, J. N. *Intermolecular and Surface Forces* 3rd edn (Elsevier, 2011).
- Ashby, P. D., Chen, L. & Lieber, C. M. Probing intermolecular forces and potentials with magnetic feedback chemical force microscopy. *J. Am. Chem. Soc.* **122**, 9467–9472 (2000).

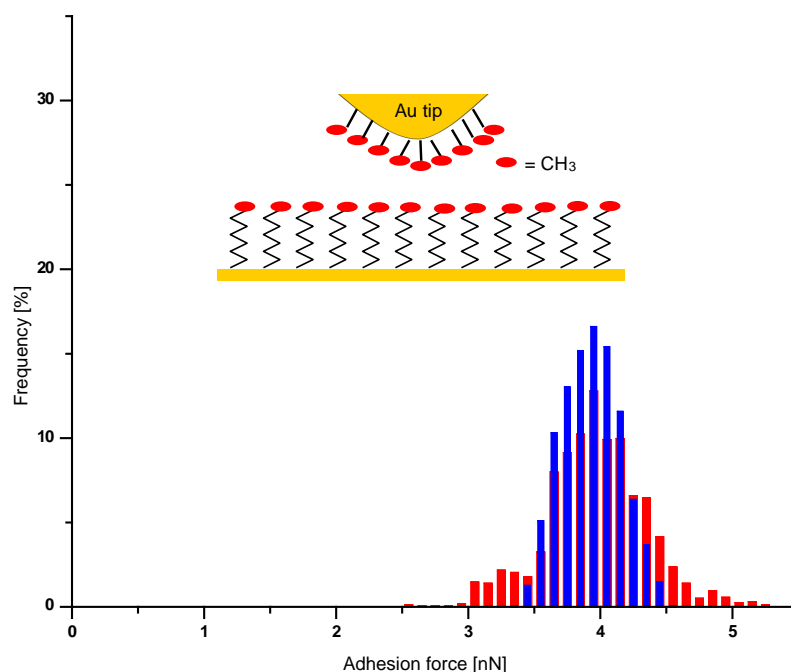
34. Seog, J. *et al.* Direct measurement of glycosaminoglycan intermolecular interactions via high-resolution force spectroscopy. *Macromolecules* **35**, 5601–5615 (2002).
35. Tian, C. S. & Shen, Y. R. Structure and charging of hydrophobic material/water interfaces studied by phase-sensitive sum-frequency vibrational spectroscopy. *Proc. Natl Acad. Sci. USA* **106**, 15148–15153 (2009).
36. Zangi, R. & Engberts, J. B. F. N. Physisorption of hydroxide ions from aqueous solution to a hydrophobic surface. *J. Am. Chem. Soc.* **127**, 2272–2276 (2005).
37. Vácha, R. *et al.* The orientation and charge of water at the hydrophobic oil droplet-water interface. *J. Am. Chem. Soc.* **133**, 10204–10210 (2011).
38. Butt, H. J., Cappella, B. & Kappl, M. Force measurements with the atomic force microscope: technique, interpretation and applications. *Surf. Sci. Rep.* **59**, 1–152 (2005).
39. Burns, A. R., Houston, J. E., Carpick, R. W. & Michalske, T. A. Molecular level friction as revealed with a novel scanning probe. *Langmuir* **15**, 2922–2930 (1999).
40. Vezenov, D. V., Noy, A. & Lieber, C. M. The effect of liquid-induced adhesion changes on the interfacial shear strength between self-assembled monolayers. *J. Adhes. Sci. Technol.* **17**, 1385–1401 (2003).
41. Cheng, R. P., Gellman, S. H. & DeGrado, W. F.  $\beta$ -Peptides: from structure to function. *Chem. Rev.* **101**, 3219–3232 (2001).
42. Pomerantz, W. C. *et al.* Lyotropic liquid crystals formed from AHC-rich beta-peptides. *J. Am. Chem. Soc.* **133**, 13604–13613 (2011).
43. Harder, P., Grunze, M., Dahint, R., Whitesides, G. M. & Laibinis, P. E. Molecular conformation in oligo(ethylene glycol)-terminated self-assembled monolayers on gold and silver surfaces determines their ability to resist protein adsorption. *J. Phys. Chem. B* **102**, 426–436 (1998).



**Extended Data Figure 1 | Influence of pH and addition of methanol (60 vol%) on adhesive interactions between self-assembled monolayers and alkyl-terminated AFM tips.** **a**, Adhesion force histograms for  $C_{10}H_{21}SH$  monolayers interacting with an alkyl-terminated AFM tip, measured as a function of pH (red, in TEA; black, in 60 vol% methanol).  $n = 3,002$  (number of test events),  $N = 3$  (number of independent samples) (TEA pH 7);  $n = 3,084$ ,  $N = 4$  (TEA pH 8);  $n = 1,076$ ,  $N = 3$  (TEA pH 9);  $n = 1,288$ ,  $N = 6$  (TEA pH 10.5);  $n = 3,812$ ,  $N = 4$  (60 vol% MeOH pH 7);  $n = 4,306$ ,  $N = 8$  (60 vol% MeOH pH 8);  $n = 1,057$ ,  $N = 4$  (60 vol% MeOH pH 9);  $n = 1,093$ ,  $N = 6$  (60 vol% MeOH pH 10.5). **b**, Histograms of adhesion forces measured between an alkyl-terminated AFM tip and monolayers formed from  $AmC_{11}H_{22}SH$ ,

reported as a function of pH (red, in TEA; blue, in 60 vol% methanol). In TEA:  $n = 1,309$ ,  $N = 5$  at pH 7;  $n = 1,797$ ,  $N = 4$  at pH 8;  $n = 1,605$ ,  $N = 4$  at pH 9;  $n = 1,009$ ,  $N = 4$  at pH 10.5. In 60 vol% methanol:  $n = 1,772$ ,  $N = 3$  at pH 7;  $n = 1,614$ ,  $N = 5$  at pH 8;  $n = 1,603$ ,  $N = 5$  at pH 9;  $n = 1,151$ ,  $N = 4$  at pH 10.5. **c**, Histograms of adhesion forces measured between an alkyl-terminated AFM tip and monolayers of  $GdmC_{11}H_{22}SH$ , measured as a function of pH (red, in TEA; green, in 60 vol% methanol). In TEA:  $n = 1,693$ ,  $N = 4$  at pH 7;  $n = 1,164$ ,  $N = 4$  at pH 8;  $n = 2,002$ ,  $N = 4$  at pH 9;  $n = 1,249$ ,  $N = 3$  at pH 10.5. In 60 vol% methanol:  $n = 1,907$ ,  $N = 4$  at pH 7;  $n = 1,211$ ,  $N = 4$  at pH 8;  $n = 2,618$ ,  $N = 5$  at pH 9;  $n = 1,178$ ,  $N = 3$  at pH 10.5. The histograms show data obtained from all pull-off force curves from all samples.

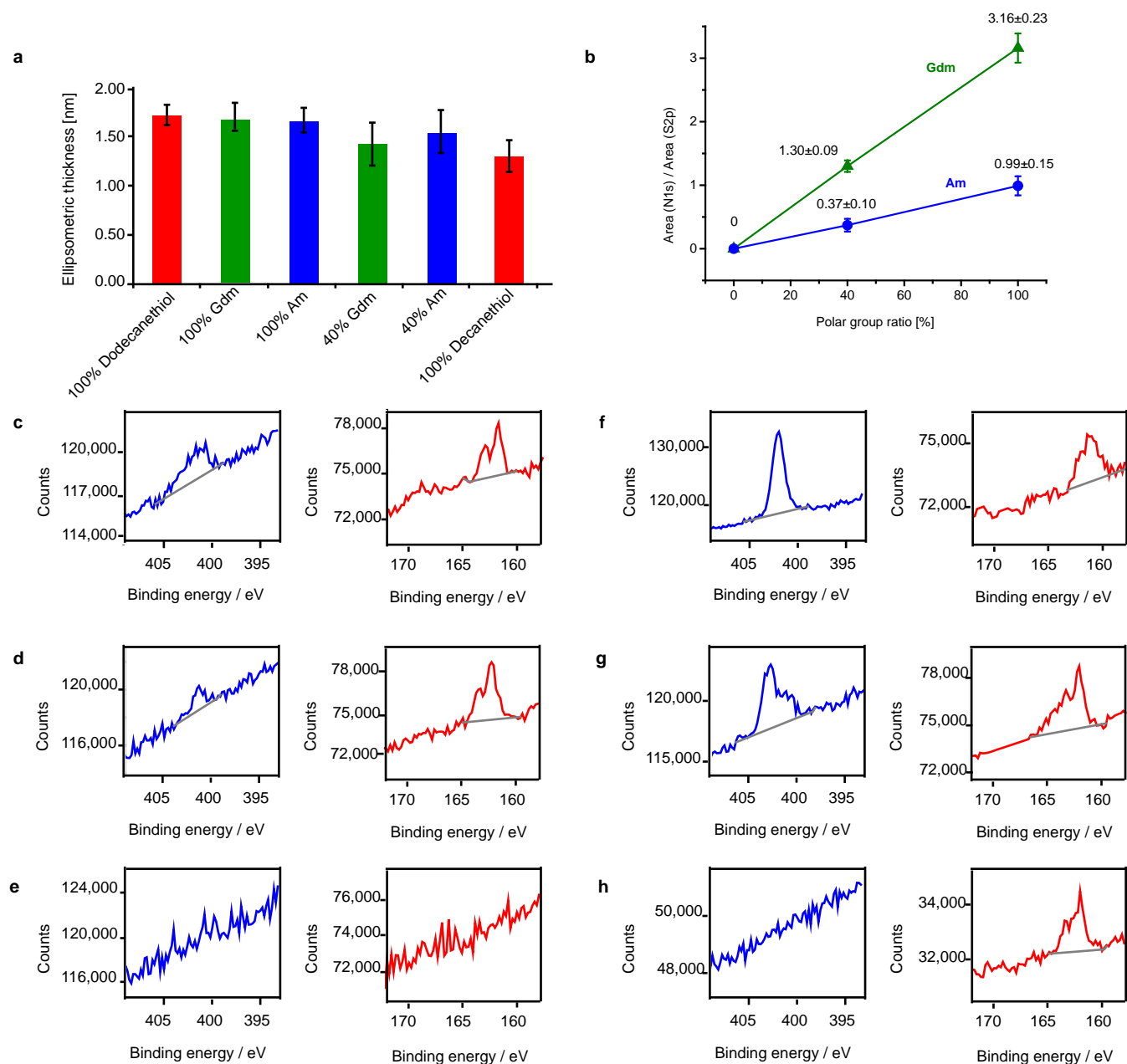




**Extended Data Figure 2 | Comparison of adhesive interactions measured between hydrophobic surfaces in pure water and in aqueous TEA.**

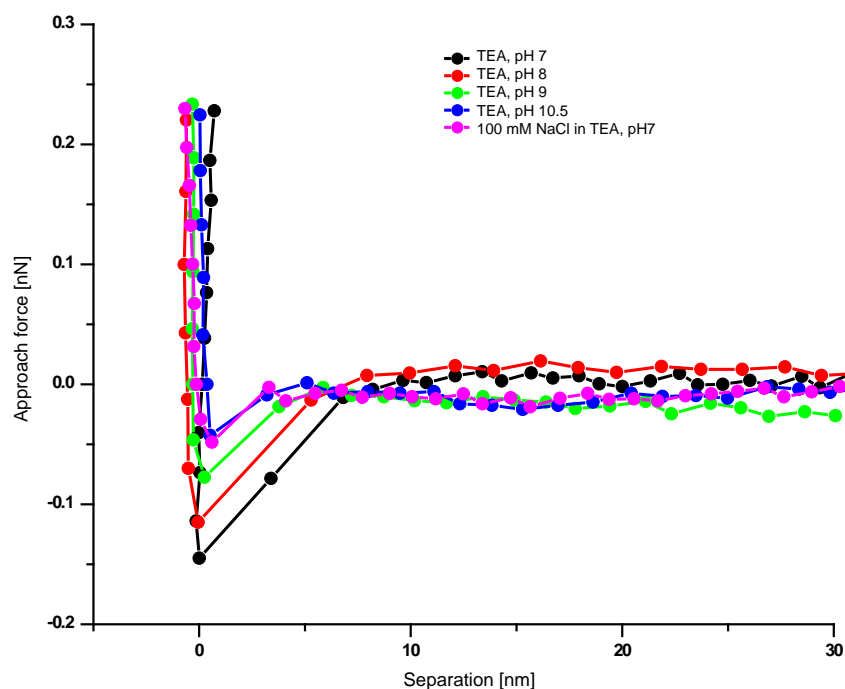
Histograms of adhesion forces for  $C_{10}H_{21}SH$  monolayers interacting with an

alkyl-terminated AFM tip under different solution conditions (red, in TEA at pH 7,  $n = 3,002$ ,  $N = 3$ ; blue, in water,  $n = 4,770$ ,  $N = 6$ ). The histograms show data obtained from all pull-off force curves from all samples.



**Extended Data Figure 3 | Characterization of the composition of mixed monolayers.** **a**, Ellipsometric thicknesses of monolayers used in this study ( $n = 3$ ,  $N = 3$ ). **b**, Ratio of nitrogen to sulphur signal, obtained by X-ray photoelectron spectroscopy ( $n = 3$ ,  $N = 3$ ), for mixed monolayers, plotted as a function of the mole fraction of the Am- or Gdm-terminated alkanethiol in the solution from which the mixed monolayers were formed. Values are means and

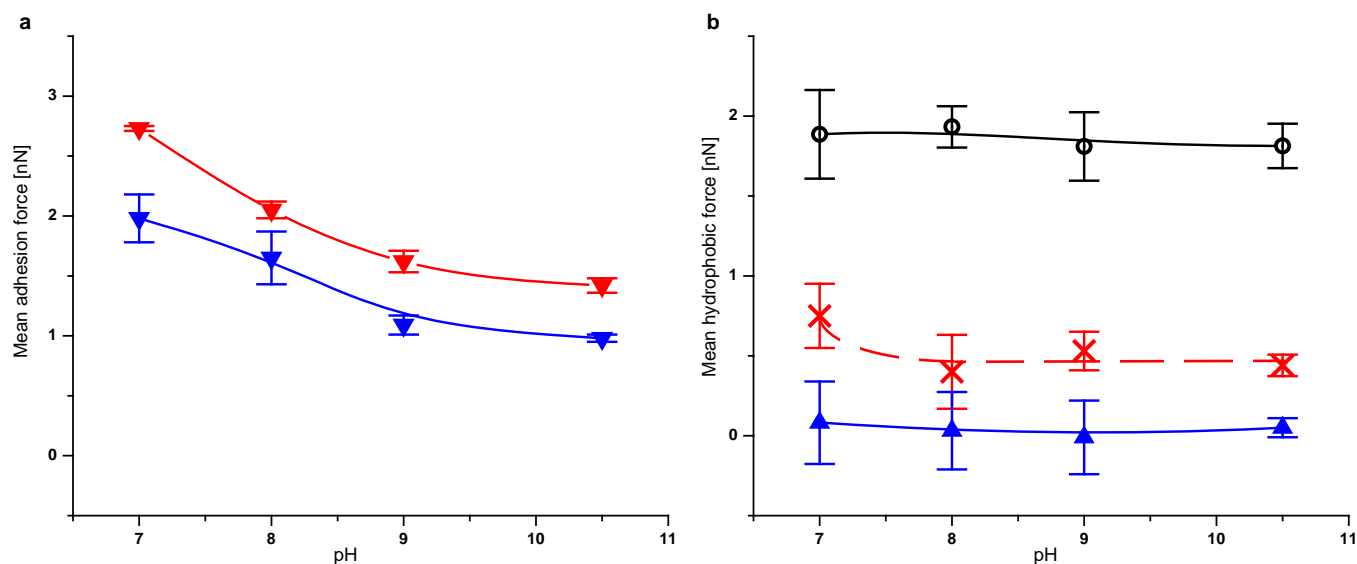
the error bars show the s.d. of three independent samples. **c–h**, Nitrogen (blue) and sulphur (red) signals obtained by X-ray photoelectron spectroscopy for mixed monolayers formed on the surfaces of gold films: GdmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH (**c**), AmC<sub>11</sub>H<sub>22</sub>SH–C<sub>10</sub>H<sub>21</sub>SH (**d**), bare gold (**e**), GdmC<sub>11</sub>H<sub>22</sub>SH (**f**), AmC<sub>11</sub>H<sub>22</sub>SH (**g**) and C<sub>10</sub>H<sub>21</sub>SH (**h**).



**Extended Data Figure 4 | Influence of pH and ionic strength on the distance dependence of the interaction of a hydrophobic AFM tip and an Am-terminated monolayer (on approach).** Approach curves for alkyl-terminated

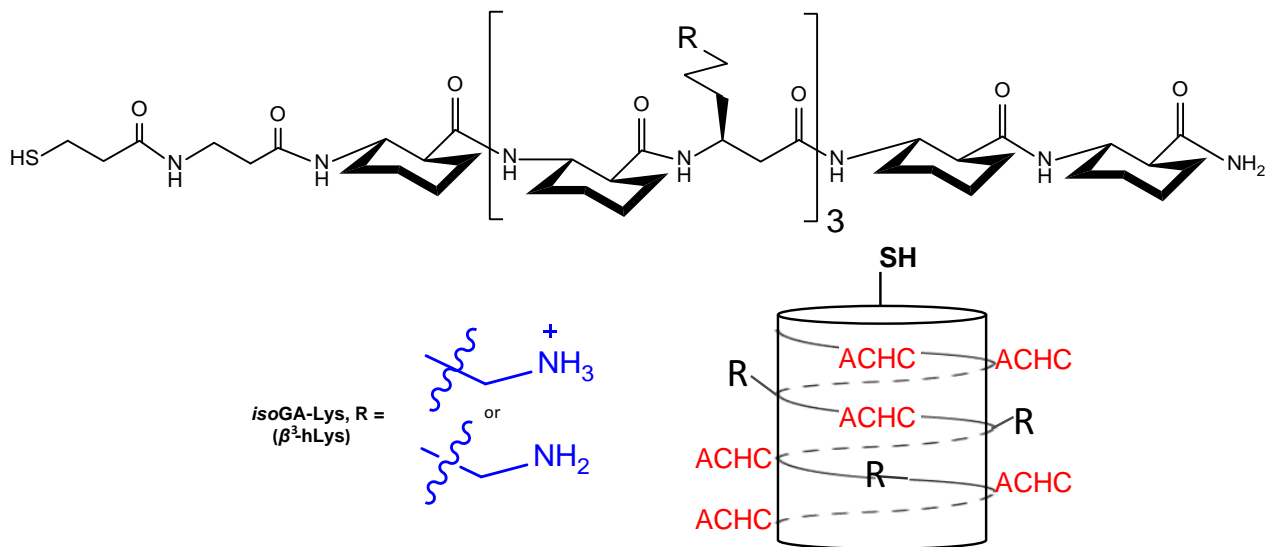
AFM tips interacting with AmC<sub>11</sub>H<sub>22</sub>SH monolayers, as measured using the indicated aqueous solution conditions.



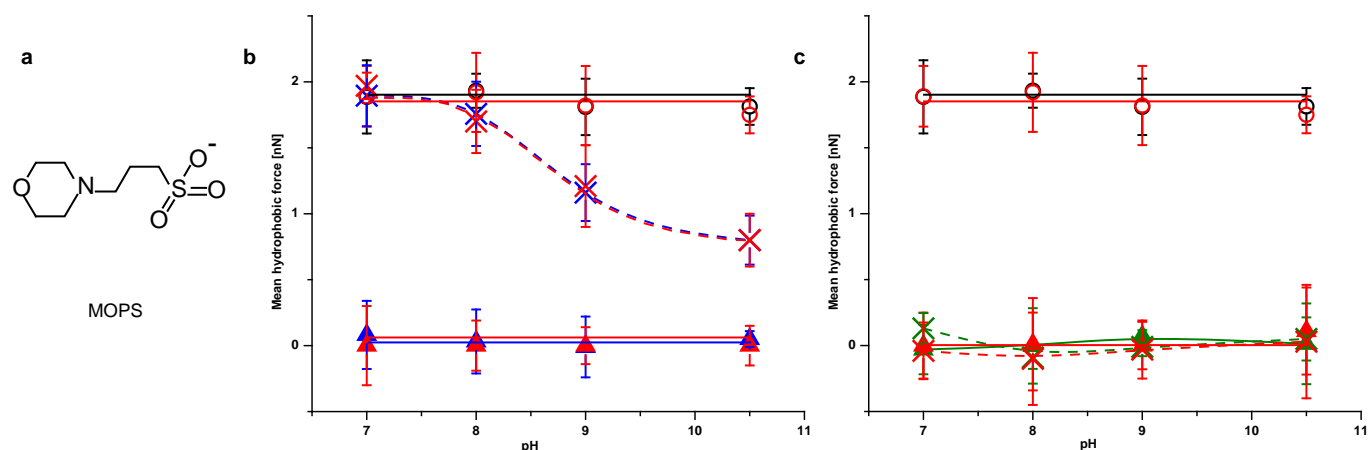


**Extended Data Figure 5 | Influence of pH and addition of methanol (60 vol%) on adhesive interaction between an alkyl-terminated AFM tip and monolayers containing 90% AmC<sub>11</sub>H<sub>22</sub>SH–10% C<sub>10</sub>H<sub>21</sub>SH.** **a**, pH dependence of mean adhesion force measured between an alkyl-terminated AFM tip and Am-containing monolayers: 90% AmC<sub>11</sub>H<sub>22</sub>SH–10% C<sub>10</sub>H<sub>21</sub>SH in either TEA (red triangles:  $n = 1,344$ ,  $N = 4$  at pH 7;  $n = 1,326$ ,  $N = 5$  at pH 8;

$n = 1,480$ ,  $N = 4$  at pH 9;  $n = 1,730$ ,  $N = 4$  at pH 10.5) or 60 vol% methanol (blue triangles:  $n = 972$ ,  $N = 4$  at pH 7;  $n = 1,548$ ,  $N = 4$  at pH 8;  $n = 1,294$ ,  $N = 4$  at pH 9;  $n = 1,176$ ,  $N = 4$  at pH 10.5). **b**, Hydrophobic contribution to the mean adhesion forces measured using 90% AmC<sub>11</sub>H<sub>22</sub>SH–10% C<sub>10</sub>H<sub>21</sub>SH (red triangles), AmC<sub>11</sub>H<sub>22</sub>SH (blue triangles) or C<sub>10</sub>H<sub>21</sub>SH (black circles) monolayers. Data show mean  $\pm$  s.e.m.



**Extended Data Figure 6 | Non-globally amphiphilic  $\beta$ -peptide.** Linear and helical representations of the non-globally amphiphilic  $\beta$ -peptide isoGA-Lys.

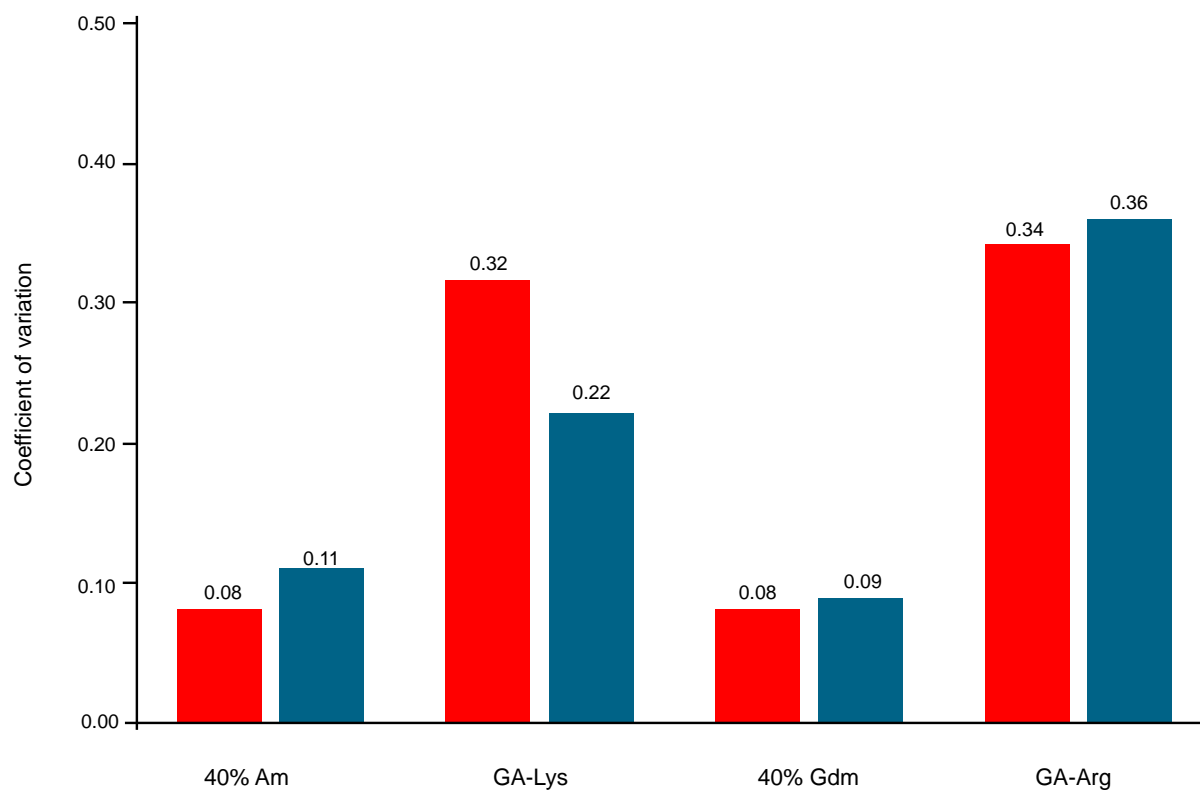


**Extended Data Figure 7 | Influence of dissolved anions (MOPS versus Cl<sup>-</sup>) on hydrophobic interaction.** **a**, The chemical structure of MOPS.

**b**, Hydrophobic contribution to the mean adhesion force measured using monolayers of AmC<sub>11</sub>H<sub>22</sub>SH-C<sub>10</sub>H<sub>21</sub>SH (red crosses, using MOPS-60 vol% methanol; blue crosses, using TEA-60 vol% methanol), AmC<sub>11</sub>H<sub>22</sub>SH (red triangles, using MOPS-60 vol% methanol; blue triangles, using TEA-60 vol% methanol) or C<sub>10</sub>H<sub>21</sub>SH (red circles, using MOPS-60 vol% methanol; black circles, using TEA-60 vol% methanol). **c**, Hydrophobic contribution to the mean adhesion force measured using monolayers of GdmC<sub>11</sub>H<sub>22</sub>SH-C<sub>10</sub>H<sub>21</sub>SH (red crosses, using MOPS-60 vol% methanol; green crosses, using TEA-60 vol% methanol), GdmC<sub>11</sub>H<sub>22</sub>SH (red triangles, using MOPS-60 vol% methanol; green triangles, using TEA-60 vol% methanol) or C<sub>10</sub>H<sub>21</sub>SH (red circles, using MOPS-60 vol% methanol; black circles, using TEA-60 vol% methanol). On C<sub>10</sub>H<sub>21</sub>SH surface:  $n = 1,702$ ,  $N = 4$  (MOPS pH 7);  $n = 1,014$ ,  $N = 4$  (MOPS pH 8);  $n = 1,006$ ,  $N = 3$  (MOPS pH 9);  $n = 1,008$ ,  $N = 4$  (MOPS pH 10.5);  $n = 1,002$ ,  $N = 4$  (MOPS-60 vol% MeOH pH 7);  $n = 1,000$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 8);  $n = 1,009$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 9);  $n = 1,001$ ,  $N = 4$  (MOPS-60 vol% MeOH pH 10.5). On AmC<sub>11</sub>H<sub>22</sub>SH-C<sub>10</sub>H<sub>21</sub>SH surface:  $n = 1,100$ ,  $N = 3$  (MOPS pH 7);  $n = 1,201$ ,  $N = 4$  (MOPS pH 8);  $n = 989$ ,  $N = 4$  (MOPS pH 9);  $n = 998$ ,  $N = 4$  (MOPS pH 10.5);

$n = 1,122$ ,  $N = 4$  (MOPS-60 vol% MeOH pH 7);  $n = 997$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 8);  $n = 1,126$ ,  $N = 4$  (MOPS-60 vol% MeOH pH 9);  $n = 1,328$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 10.5). On GdmC<sub>11</sub>H<sub>22</sub>SH surface:  $n = 1,000$ ,  $N = 3$  (MOPS pH 7);  $n = 1,001$ ,  $N = 3$  (MOPS pH 8);  $n = 1,002$ ,  $N = 3$  (MOPS pH 9);  $n = 1,001$ ,  $N = 3$  (MOPS pH 10.5);  $n = 1,003$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 7);  $n = 1,000$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 8);  $n = 1,001$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 9);  $n = 1,005$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 10.5). On GdmC<sub>11</sub>H<sub>22</sub>SH-C<sub>10</sub>H<sub>21</sub>SH surface:  $n = 999$ ,  $N = 3$  (MOPS pH 7);  $n = 1,001$ ,  $N = 3$  (MOPS pH 8);  $n = 1,000$ ,  $N = 3$  (MOPS pH 9);  $n = 999$ ,  $N = 3$  (MOPS pH 10.5);  $n = 1,002$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 7);  $n = 999$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 8);  $n = 1,001$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 9);  $n = 1,002$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 10.5). On AmC<sub>11</sub>H<sub>22</sub>SH surface:  $n = 1,004$ ,  $N = 3$  (MOPS pH 7);  $n = 1,002$ ,  $N = 3$  (MOPS pH 8);  $n = 1,002$ ,  $N = 3$  (MOPS pH 9);  $n = 1,002$ ,  $N = 3$  (MOPS pH 10.5);  $n = 1,001$ ,  $N = 3$  (MOPS 60 vol% MeOH pH 7);  $n = 1,001$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 8);  $n = 1,001$ ,  $N = 3$  (MOPS-60 vol% MeOH pH 9);  $n = 1,001$ ,  $N = 4$  (MOPS-60 vol% MeOH pH 10.5). Measurements were conducted as described in Methods. Data show mean  $\pm$  s.e.m. Lines are drawn to guide the eye.





**Extended Data Figure 8 | Characterization of the widths of adhesion force histograms.** The coefficient of variation was calculated from histograms of the adhesion forces measured using the indicated surfaces (red, in TEA at

pH 9; blue, in 60 vol% methanol). Measurements were conducted as detailed in Methods.

Extended Data Table 1 | Statistical information for data shown in Fig. 2b related to the influence of methanol

Vol% Methanol	The number of test events / <i>n</i>	The number of independent samples / <i>N</i>
0%	3,002	3
20%	2,128	4
40%	1,034	3
60%	3,810	4
80%	1,053	3
100%	1,004	3

Extended Data Table 2 | Statistical information for data shown in Fig. 2b (pH-dependent data) and Figs 3 and 4

Figure	Monolayer	The number of test events / The number of samples (n / N)							
		TEA pH7	TEA pH8	TEA pH9	TEA pH10.5	60% MeOH pH7	60% MeOH pH8	60% MeOH pH9	60% MeOH pH10.5
2b	C <sub>10</sub> H <sub>21</sub> SH	3,002/ 3	3,084/ 4	1,076 / 3	1,288 / 6	3,812/ 4	4,306/ 8	1,057/ 4	1,093/ 6
3a	AmC <sub>11</sub> H <sub>22</sub> SH	1,309/ 5	1,797/ 4	1,605/ 4	1,009/ 4	1,772/ 3	1,614/ 5	1,603/ 5	1,151/ 4
3a	40% AmC <sub>11</sub> H <sub>22</sub> SH/ 60% C <sub>10</sub> H <sub>21</sub> SH	1,709/ 4	1,800/ 3	2,085/ 5	922/ 3	1,626/ 4	1,569/ 4	2,982/ 7	1,669/ 4
3d	GdmC <sub>11</sub> H <sub>22</sub> SH	1,693/ 4	1,164/ 4	2,002/ 4	1,249/ 3	1,907/ 4	1,211/ 4	2,618/ 5	1,178/ 3
3d	40% GdmC <sub>11</sub> H <sub>22</sub> SH/ 60% C <sub>10</sub> H <sub>21</sub> SH	3,132/ 6	2,997/ 5	3,218/ 6	2,949/ 6	3,030/ 6	3,109/ 5	3,018/ 6	2,912/ 6
4 a-d	GA-Lys	3,175/ 4	1,957/ 4	1,606/ 4	1,654/ 4	3,666/ 4	3,268/ 4	1,542/ 4	1,867/ 4
4 e-h	GA-Arg	3,676/ 6	4,221/ 7	3,573/ 6	2,580/ 5	2,191/ 6	3,785/ 6	5,004/ 5	2,582/ 5



# Non-stabilized nucleophiles in Cu-catalysed dynamic kinetic asymmetric allylic alkylation

Hengzhi You<sup>1</sup>, Emeline Rideau<sup>1</sup>, Mireia Sidera<sup>1</sup> & Stephen P. Fletcher<sup>1</sup>

The development of new reactions forming asymmetric carbon–carbon bonds has enabled chemists to synthesize a broad range of important carbon-containing molecules, including pharmaceutical agents, fragrances and polymers<sup>1</sup>. Most strategies to obtain enantiomerically enriched molecules rely on either generating new stereogenic centres from prochiral substrates or resolving racemic mixtures of enantiomers. An alternative strategy—dynamic kinetic asymmetric transformation—involves the transformation of a racemic starting material into a single enantiomer product, with greater than 50 per cent maximum yield<sup>2,3</sup>. The use of stabilized nucleophiles ( $pK_a < 25$ , where  $K_a$  is the acid dissociation constant) in palladium-catalysed asymmetric allylic alkylation reactions has proved to be extremely versatile in these processes<sup>4,5</sup>. Conversely, the use of non-stabilized nucleophiles in such reactions is difficult and remains a key challenge<sup>6–9</sup>. Here we report a copper-catalysed dynamic kinetic asymmetric transformation using racemic substrates and alkyl nucleophiles. These nucleophiles have a  $pK_a$  of  $\geq 50$ , more than 25 orders of magnitude more basic than the nucleophiles that are typically used in such transformations. Organometallic reagents are generated *in situ* from alkenes by hydrometallation and give highly enantioenriched products under mild reaction conditions. The method is used to synthesize natural products that possess activity against tuberculosis and leprosy, and an inhibitor of *para*-aminobenzoate biosynthesis. Mechanistic studies indicate that the reaction proceeds through a rapidly isomerizing intermediate. We anticipate that this approach will be a valuable complement to existing asymmetric catalytic methods.

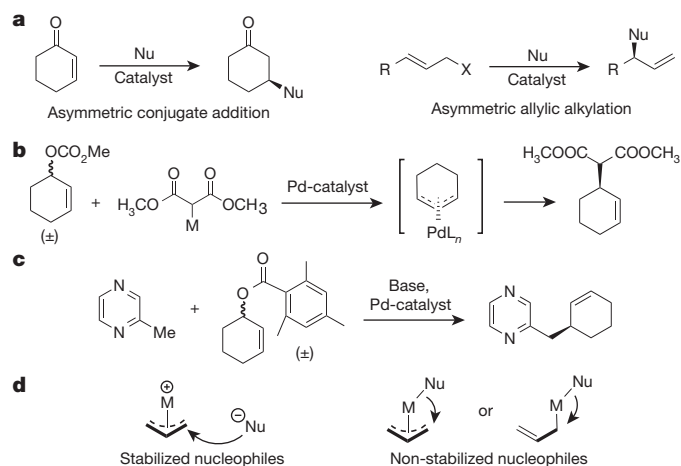
Two strategies for generating single enantiomer compounds using enantioselective catalysis have become widely used: the first strategy generates a chiral product by a selective reaction which introduces asymmetry to a prochiral substrate (Fig. 1a). This prochiral approach has proven important in the development of catalytic asymmetric reactions, despite the rather limited availability of prochiral substrates when compared to chiral substrates. The second widely used strategy is to start from a racemic mixture of chiral starting materials. Here the catalyst selectively reacts with one of the two enantiomers allowing differentiation (or resolution) of the enantiomers, but the yield is necessarily limited to 50%, as the undesired enantiomer remains as starting material. An efficient variation of this second strategy is to couple enantiomer differentiation with interconversion of the enantiomers, commonly called dynamic kinetic resolution. This strategy allows yields of more than 50% from chiral starting materials<sup>2,3</sup>.

Transition-metal-catalysed asymmetric allylic alkylation (AAA) reactions have proven to be powerful tools for creating new carbon–carbon bonds<sup>4,9–11</sup>. Twenty years ago, palladium-catalysed AAA reactions were reported in which racemic mixtures were converted to single enantiomer products with high yield and enantioselectivity (Fig. 1b)<sup>12</sup>. The use of stabilized nucleophiles ( $pK_a < 25$ ) in metal-catalysed AAA reactions is now well established in catalysis and synthesis<sup>5,9</sup>. These reactions are dynamic kinetic asymmetric transformations (DYKATs) as they form one enantiomer of a new product from both enantiomers of a racemic starting material, and no ‘resolved’ starting material is recovered<sup>3,5</sup>. Mechanistic studies reveal that Pd-catalysed DYKATs operate through a

pathway where both enantiomers of the starting material are converted to a common pseudo-prochiral intermediate, and a subsequent enantioselective step creates the new stereogenic centre<sup>6</sup>. The use of non-stabilized nucleophiles in such reactions is a long-standing research goal<sup>6–9</sup> and efforts to broaden the scope of DYKAT to non-stabilized carbon nucleophiles now allows certain specific ‘less-stabilized’ partners to be used in Pd-catalysed asymmetric (for example, Fig. 1c)<sup>7,13</sup> and non-enantioselective procedures<sup>8</sup>.

One difficulty in expanding the scope of DYKATs is that metal-catalysed AAAs do not rely on a single mechanistic pathway<sup>4</sup>. In particular, there is a significant mechanistic distinction between stabilized and non-stabilized nucleophiles (Fig. 1d)<sup>4,6,7,10</sup>. In the case of stabilized nucleophiles, the key bond-making event occurs outside of the coordination sphere of the metal, so that the stereochemistry is determined when the nucleophile attacks a carbon atom of a  $\pi$ -allyl-Pd intermediate<sup>14</sup>. In contrast, when using non-stabilized nucleophiles bond formation probably occurs through reductive elimination of an intermediate in which the nucleophile is bound to the metal centre<sup>15</sup>.

Catalytic asymmetric reactions with non-stabilized nucleophiles often use copper catalysts<sup>16,17</sup>. These reactions have proven much more difficult to study mechanistically than their Pd-catalysed counterparts, as non-stabilized organometallic reagents are typically very reactive, hindering the isolation of intermediates and characterization of pathways. Copper-catalysed allylic alkylations are generally accepted to operate via



**Figure 1 | Asymmetric allylic alkylation (AAA) procedures.** **a**, Typical examples of the prochiral approach to asymmetric catalysis, in which a species selectively adds to one face of a prochiral substrate, exemplified (on the left) by asymmetric conjugate addition and (on the right) by AAA. Nu, nucleophile; R, generic substituent (or generic group); X, leaving group. **b**, A prototypical Pd-catalysed DYKAT, where AAA with a stabilized carbon nucleophile transforms a racemic mixture of starting materials into a single highly enantioenriched product. **c**, An example of DYKAT in AAA using heterocyclic nucleophiles with a  $pK_a > 25$ . **d**, Transition-metal-catalysed allylic substitution mechanisms with stabilized and non-stabilized nucleophiles.

<sup>1</sup>Department of Chemistry, Chemistry Research Laboratory, University of Oxford, 12 Mansfield Road, Oxford OX1 3TA, UK.

regiospecific *anti*-addition mechanisms, where selective rate-determining  $S_N2'$  oxidative additions give copper(III) intermediates, and fast reductive eliminations form the final product<sup>16,17</sup>. A key distinction between Cu- and Pd-AAA reactions is a lack of an efficient  $\sigma$ - to  $\pi$ -isomerization<sup>18</sup> in the case of copper, so that if reductive elimination is fast, a  $\pi$ -allyl-Cu species will not form and reactions involving racemic starting materials will lead to racemic products.

Copper-catalysed AAA on prochiral materials, where selective oxidative addition occurs on one enantiotopic face of the substrate, can be readily achieved<sup>10,16–18</sup>. However, the generally accepted Cu-catalysed mechanisms would seem to preclude efficient asymmetric reactions with chiral racemic substrates, where addition would be controlled by the stereogenic centre of the substrate, not the catalyst<sup>18</sup>. We note that racemization of enantiopure starting materials during Cu-catalysed allylic alkylation has been reported, suggesting the possibility of generating useful  $\pi$ -allyl-Cu species and DYKATs<sup>19</sup>. Building on this work, promising initial results have been obtained with Grignard reagents<sup>20</sup>, but despite extensive examination, and the elucidation of interesting mechanistic pathways, generally useful procedures have not emerged<sup>21,22</sup>.

Based on the observation that alkylzirconium reagents undergo highly enantioselective conjugate addition at ambient temperature<sup>23,24</sup>, whereas procedures with traditional non-stabilized carbon nucleophiles (based on Mg, Al, and Zn species) are normally performed at very low temperatures<sup>16,17</sup>, we hypothesized that alkylzirconium reagents may allow efficient AAA from racemic substrates. Here we describe AAA of racemic cyclic allylchlorides with alkylzirconocenes initiated by copper catalysts. Highly enantioenriched products are obtained at room temperature in convenient procedures, in which all of the reaction components are commercially available.

We began by exploring the coupling of racemic 3-substituted cyclohexenes **1** and the alkylzirconium species generated *in situ* from 4-phenyl-1-butene (**2** in Table 1) and the Schwartz reagent ( $Cp_2ZrHCl$ ) under a

variety of reaction conditions. The use of Pd- and Ir-based catalysts did not lead to product **3a**, but when Cu-catalysts were used, product **3a** was obtained with varying degrees of enantioselectivity. The combination of 3-chloro-cyclohexene **1d**, phosphoramidite **C**<sup>11</sup>, and  $(tBuCN)_2Cu\cdot OTf$  in dichloromethane provided **3a** in quantitative yield with 71% enantiomeric excess (e.e.; Table 1, entry 6). Variation of the copper counterion under these conditions showed that CuI gave the highest e.e. (89% e.e., entry 10) and chloroform gave the best results of the solvents examined (93% e.e., entry 11). Varying the temperature (entries 12 and 13) showed that slightly higher enantioselectivity was obtained by cooling the reaction to 0 °C (93% versus 95% e.e.) but we chose to perform the rest of our studies at room temperature because of operational convenience and under these conditions full conversion was always observed overnight. Examination of bromide **1a** under these conditions (entry 14) showed that the less reactive chloride **1c** gave superior results.

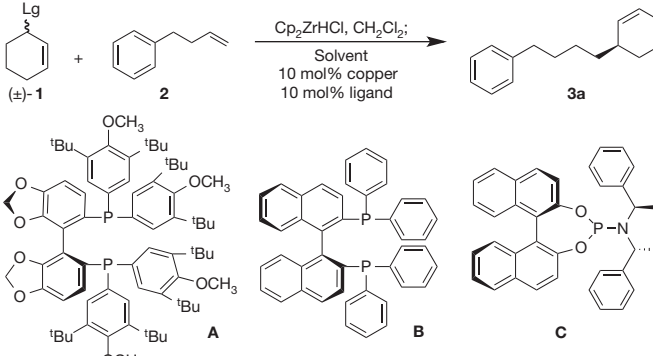
The scope of the alkyl coupling partners was investigated using a variety of simple gaseous alkenes (giving products **3b–d**) and unfunctionalized primary alkenes (**3e–g**) (Fig. 2). High yields and consistently high enantioselectivities were obtained. Functionalized alkenes (**3h–3r**) were also examined: halogens, alkynes, aromatic rings, ethers, protected alcohols, trifluoromethyl groups and preformed stereogenic centres were all found to be compatible with the procedure. Electron rich alkenes such as allylsilane (to give **3i**) and styrenes (to give **3m–o**) also work well. We also examined the use of 5-membered rings with this procedure; we found that it is suitable for the direct preparation of highly enantioenriched cyclopentenones (**4a–c**) from an easily prepared racemic 5-membered allylchloride.

To probe practical aspects of this chemistry, scale-up reactions were performed (Fig. 3a). Despite excellent conversion (>90%), the yield of **4b** on a 15 mmol scale was low because the product and the residual alkene have similar polarities and boiling points, complicating isolation. Even so, we were easily able to obtain more than 1.8 g of pure **4b** (58% yield, 94% e.e.). When 1-hexene was used (to give **4a**), we arbitrarily reduced the ratio of alkene and Schwartz reagent used (to 2 and 1.8 equiv. respectively) and observed improved yield (86% yield, 91% e.e., 1.2 g) relative to using our standard test conditions (Fig. 2). Functionalized **3n** containing a bromide (Fig. 3b) was also synthesized on a preparative scale (88% yield, 87% e.e., 0.75 g) and a subsequent Heck reaction to provide *cis*-fused tricycle **5** (88% yield) demonstrates that these reactions can be used to rapidly access complex molecular frameworks.

To demonstrate that this method allows rapid access to important structural motifs, we performed asymmetric syntheses of biologically active cyclopentene-containing natural products (Fig. 3c). This approach provides short, straightforward, flexible syntheses which compare favourably to those previously reported<sup>25</sup>, and will facilitate the study of structural analogues. Hydnocarpic acid (**6**) and chaulmoogric acid (**7**) are leprosy treatments used in ancient and traditional medicine: in both cases, the cyclopentenyl ring is a requirement for biological activity. Hydnocarpic acid is believed to act by blocking the activity of biotin or inhibiting microbial biotin synthesis<sup>26</sup>, while the activity of chaulmoogric acid is probably due to incorporation into the cell wall lipids of *Mycobacterium leprae*<sup>27</sup>. Recent reports show that chaulmoogric acid and anthelmintic C (**8**) significantly inhibit *Mycobacterium tuberculosis* growth with minimum inhibitory concentration values of 9.82 and 4.38  $\mu$ M, respectively<sup>28</sup>. Anthelmintic C has also recently been found to inhibit *para*-aminobenzoic acid biosynthesis (*p*ABAB)<sup>28</sup>; inhibitors of *p*ABAB are important leads for the development of new antibiotics, as this pathway is not found in humans<sup>28</sup>.

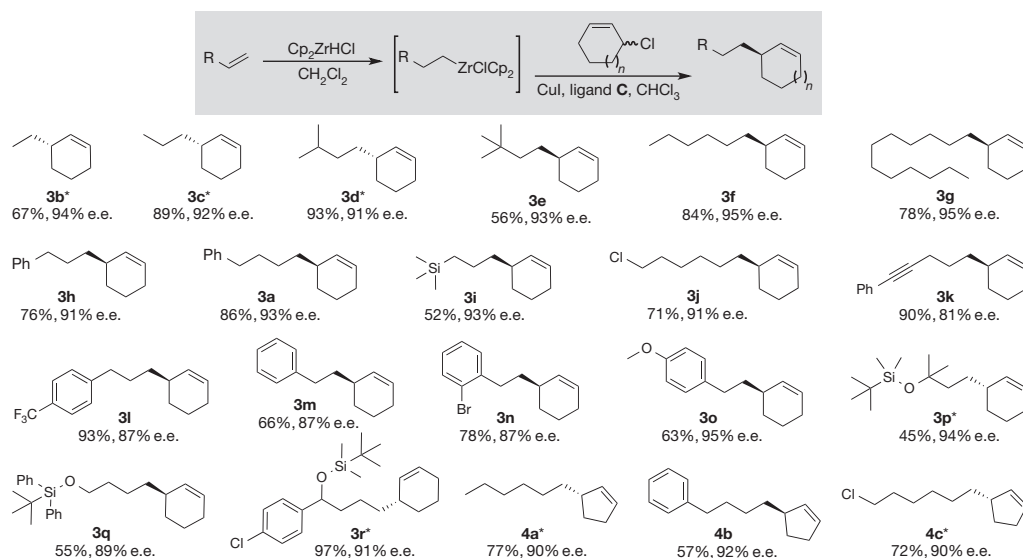
On the basis of previously reported reactions it seemed likely that DYKAT was occurring by one of the four following pathways: (1) the reaction proceeds through a pseudo-*meso*- or prochiral  $\pi$ -allyl-Cu intermediate<sup>19,20</sup> similar to the Pd- or Ir-intermediates observed with stabilized nucleophiles<sup>6</sup>; (2) a related scenario, in which the reaction proceeds through rapidly interconverting intermediates derived from **1d** (such as  $\sigma$ -allyl-Cu species), one enantiomer of which undergoes a selective reaction<sup>5</sup>; (3) the two enantiomers of **1d** undergo two different

**Table 1 | Selected optimization experiments**



Entry	Leaving group	Ligand	Copper source	Solvent	Temp. (°C)	e.e. (%)
1	Br ( <b>1a</b> )	<b>A</b>	$(tBuCN)_2Cu\cdot OTf$	$CH_2Cl_2$	25	4
2	OP(O)(OEt) <sub>2</sub> ( <b>1b</b> )	<b>A</b>	$(tBuCN)_2Cu\cdot OTf$	$CH_2Cl_2$	25	10
3	O <sub>2</sub> CCF <sub>3</sub> ( <b>1c</b> )	<b>A</b>	$(tBuCN)_2Cu\cdot OTf$	$CH_2Cl_2$	25	10
4	Cl ( <b>1d</b> )	<b>A</b>	$(tBuCN)_2Cu\cdot OTf$	$CH_2Cl_2$	25	35
5	Cl	<b>B</b>	$(tBuCN)_2Cu\cdot OTf$	$CH_2Cl_2$	25	1
6	Cl	<b>C</b>	$(tBuCN)_2Cu\cdot OTf$	$CH_2Cl_2$	25	71
7	Cl	<b>C</b>	CuCl/AgSbF <sub>6</sub>	$CH_2Cl_2$	25	51
8	Cl	<b>C</b>	CuCl	$CH_2Cl_2$	25	49
9	Cl	<b>C</b>	CuBr	$CH_2Cl_2$	25	73
10	Cl	<b>C</b>	CuI	$CH_2Cl_2$	25	89
11	Cl	<b>C</b>	CuI	$CHCl_3$	25	93
12	Cl	<b>C</b>	CuI	$CHCl_3$	0	95
13	Cl	<b>C</b>	CuI	$CHCl_3$	60	77
14	Br ( <b>1a</b> )	<b>C</b>	CuI	$CHCl_3$	25	77

Optimization was performed for the AAA of racemic materials bearing different leaving groups and 4-phenyl-1-butene, which was hydrometallated using the Schwartz reagent. Different copper sources, ligands, solvents and temperatures were examined. The enantiomeric excess (e.e.) was determined by HPLC. Full conversion was observed in all cases shown.



**Figure 2 | Dynamic kinetic AAA with alkylzirconium nucleophiles generated from alkenes.** Top row, alkenes react with the Schwartz reagent ( $\text{Cp}_2\text{ZrHCl}$ ) in  $\text{CH}_2\text{Cl}_2$  to generate alkylzirconocene species *in situ* which undergo AAA with racemic cyclic allylic chlorides. Cp, cyclopentadienyl. Conditions: 2.5 equiv. alkene, 2 equiv.  $\text{Cp}_2\text{ZrHCl}$ , 10 mol%  $\text{CuI}$ , 10 mol% ligand **C**, room temperature overnight, argon atmosphere. Lower rows, yield

and enantiomeric excess of illustrated products. Yield refers to pure isolated compounds; the enantiomeric excess was either determined directly by high-performance liquid chromatography, or on the corresponding epoxides by gas chromatography, using a chiral non-racemic stationary phase.

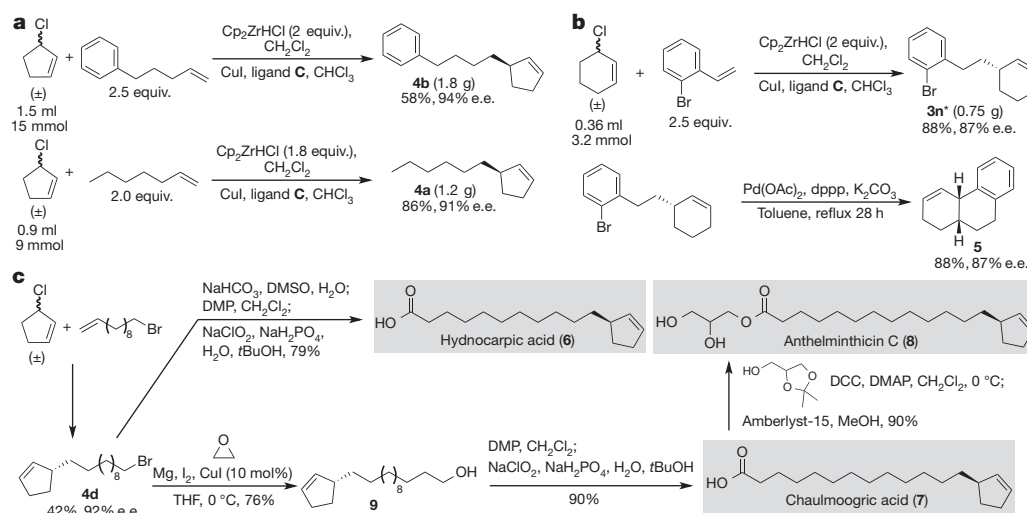
\*Prepared using (*S,S,S*)-**C** (the enantiomer of the ligand shown in Table 1).

reactions, both giving the same product (an enantio-convergent transformation)<sup>21</sup>; or (4) **1d** racemizes during the reaction<sup>2</sup> and AAA is selective for one of the two enantiomers.

To gain insight into the mechanism, we followed reactions *in situ* using  $^1\text{H}$  NMR spectroscopy. During these experiments, we observed clean conversion of **1d** to either **3a** or **3f**, and the formation of low, but fairly constant, concentrations of allyliodide **10** (Fig. 4a, also see Supplementary Information). The initial drop in the concentration of **1d** is an artefact of inconsistent NMR field inhomogeneities ('shimming') at early stages of the reaction. We were unable to observe the formation of alkyl-copper species (that would be produced by transmetalation from zirconium to copper) or  $\sigma$ - or  $\pi$ -allyl-copper species under these

conditions, and we are unable to rule out the possibility of such species playing a role in the enantioselectivity of these reactions.

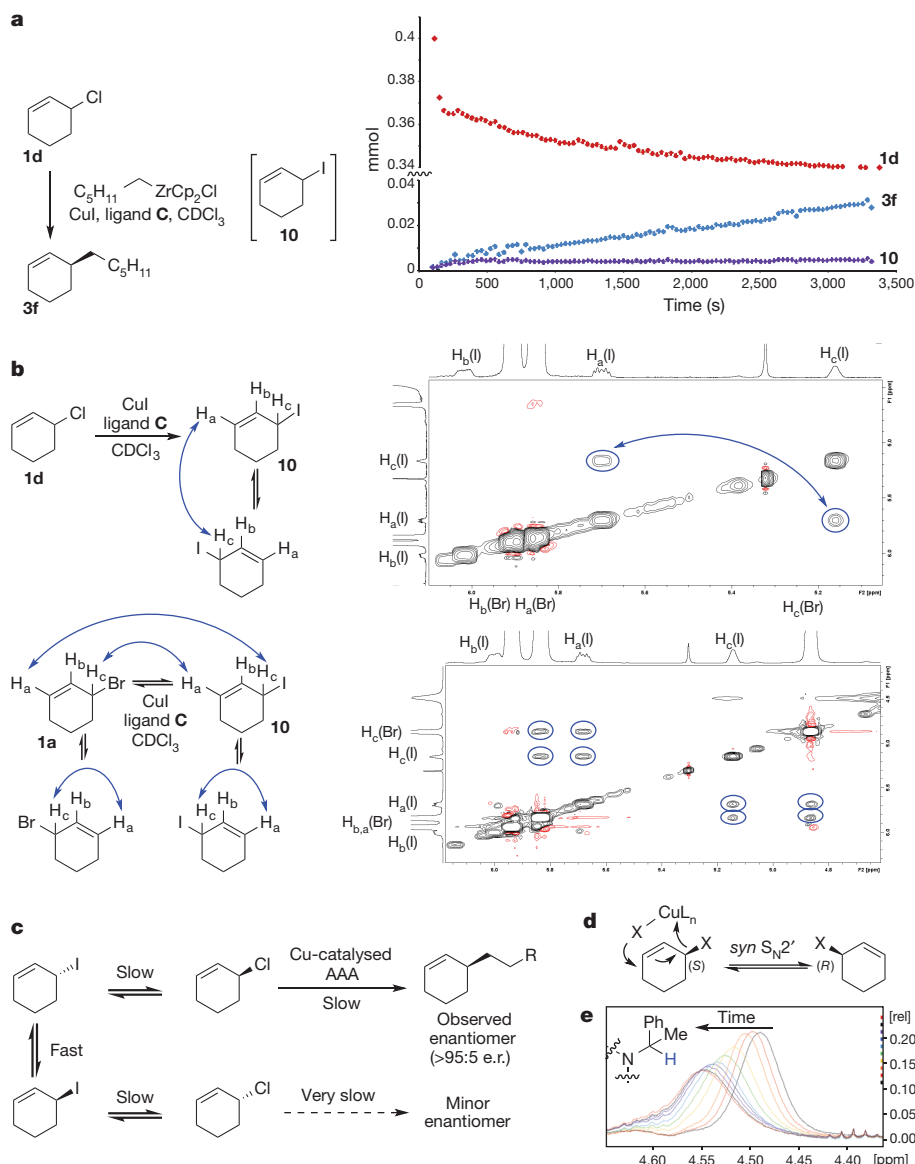
Examination of allyliodide **10** generated from **1d** (by  $\text{CuI}$  and **C** in  $\text{CDCl}_3$ ) using 2-dimensional nuclear Overhauser effect NMR spectroscopy (NOESY) shows that the  $\text{H}_a$  and  $\text{H}_c$  protons of **10** rapidly interconvert (as indicated by cross-peaks between these proton signals, Fig. 4b, top right). At ambient temperature **1d** did not undergo observable dynamic processes, but at 55 °C (see Supplementary Information) we were able to detect interconversion between **1d** and **10**, demonstrating that the allyliodide and allylchloride interconvert, but not fast enough to be detected on the NMR timescale at room temperature. Under these conditions **1d** was not observed to isomerize directly. Monitoring an



**Figure 3 | Scale-up and applications of the method to access tricyclic structures and natural products.** a, Experiments on a preparative scale to give >1 g of cyclopentene products. b, Preparative-scale reaction on a 6-membered ring followed by Heck reaction to give *cis*-fused tricycle **5**. c, Concise asymmetric synthesis of hydnocarpic acid (**6**), chaulmoogric acid (**7**) and anthelminticin C (**8**). The syntheses start from commercially available 11-bromo-1-undecene to give bromo-substituted **4d** which is converted to

**6** (by hydrolysis, then oxidation), or to **9** (by epoxide opening chain extension) followed by oxidation to **7**. Coupling **7** with racemic solketal, then acid catalysed deprotection, gave anthelminticin C (**8**). \*Made using (*S,S,S*)-**C** (the enantiomer of the ligand shown in Table 1). DMP, Dess-Martin periodinane; DCC, dicyclohexylcarbodiimide; dppp, 1,3-bis(diphenylphosphino)propane; DMAP, 4-dimethylaminopyridine. Amberlyst 15 is a polymeric acid resin. See Supplementary Information for full details.





**Figure 4 | Mechanistic analysis.** **a**, Reaction kinetics as monitored by *in situ*  $^1\text{H}$  NMR spectroscopy. Formation of an allyliodide intermediate (**10**) is observed. Plot at right shows concentration versus time. **b**, Examination of intermediates **10** (generated from **1d** and **C** in  $\text{CDCl}_3$ ) by NOESY: when generated from **1d** (top, proton chemical shifts on both axes, shift ranges as follows: vertical, 4.5–6.3 p.p.m.; horizontal, 6.2–5.05 p.p.m.), allyliodide protons labelled  $\text{H}_a$  and  $\text{H}_c$  exchange. When generated from **1a** (bottom, proton chemical shifts on both axes, shift ranges as follows: vertical, 4.1–6.3 p.p.m.; horizontal, 6.2–4.6 p.p.m.), rapid  $\text{H}_a/\text{H}_c$  isomerization is observed within both **1a** and **10**. Rapid exchange between **10** and **1a** is also observed.

AAA reaction to form **3a** from **1d** by *in situ* NOESY (see Supplementary Information) demonstrates that allyliodide **10** undergoes dynamic isomerization during the AAA itself.

When allylbromide **1a** is dissolved in  $\text{CDCl}_3$  at room temperature, no dynamic processes are observed by NOESY (see Supplementary Information), but in the presence of **CuI** and **C** a variety of fast exchanges are observed (Fig. 4b, bottom right). Allyliodide **10** is formed, and both **1a** and **10** rapidly isomerize (as indicated by cross peaks between  $\text{H}_a(\text{Br})/\text{H}_c(\text{Br})$  of **1a** and  $\text{H}_a(\text{I})/\text{H}_c(\text{I})$  of **10**). Rapid interconversion of **1a** and **10** occurs as indicated by cross peaks between  $\text{H}_a(\text{Br})/\text{H}_c(\text{I})$  and  $\text{H}_a(\text{I})/\text{H}_c(\text{Br})$ . As cross peaks between  $\text{H}_a(\text{Br})/\text{H}_a(\text{I})$  and  $\text{H}_c(\text{Br})/\text{H}_c(\text{I})$  are not observed, interconversion appears to exclusively occur by  $\text{S}_{\text{N}}2'$  mechanisms, although  $\text{S}_{\text{N}}2$  exchange may occur through minor pathways not observable on the NMR timescale.

Cross peaks between  $\text{H}_a(\text{Br})/\text{H}_c(\text{I})$  and  $\text{H}_a(\text{I})/\text{H}_c(\text{Br})$ , but not between  $\text{H}_a(\text{Br})/\text{H}_a(\text{I})$  or  $\text{H}_c(\text{Br})/\text{H}_c(\text{I})$ , suggest interconversion occurs by an  $\text{S}_{\text{N}}2'$  pathway. **c**, Proposed mechanism for the DYKAT: as **1d** racemizes via **10**, one enantiomer of **1d** undergoes selective AAA. **d**, Possible racemization mechanism: metal-assisted *syn-SN2'* attack. **e**,  $^1\text{H}$  chemical shifts (vertical, relative intensity, 0.00–0.25; horizontal, chemical shift, 4.65–4.00 p.p.m.) of the benzylic protons of **C** are consistent with a change in metal-ligand aggregation, as a function of time (colours are for different spectra, taken at different times), from the initial  $[\text{L}_n\text{CuI}]_n$  species.

Taken together, these studies point to a mechanism (Fig. 4c) where **1d** racemizes, via **10**, and selective AAA of one of the enantiomers of **1d** gives a product with high enantiomeric excess. When using copper halides, higher enantioselectivities are obtained when the halide is a better nucleophile and leaving group (that is,  $\text{I} > \text{Br} > \text{Cl}$ , Table 1, entries 8–10), facilitating interconversion, but the overall reaction rates with **CuCl** and **CuI** are similar, suggesting that it is **1d** that undergoes AAA.

Racemization probably occurs through *syn-SN2'* displacement reactions:  $\text{S}_{\text{N}}2'$  pathways are observed by NOESY, and *anti-SN2'* reactions would not change the absolute sense of stereochemistry of the allylic halides. Many allylic substitutions are *syn-SN2'* selective, and 'ion-pair' nucleophiles undergo selective *syn-SN2'* reactions through cyclic transition structures<sup>29</sup>. Our studies are consistent with the copper-mediated *syn-SN2'* pathway shown in Fig. 4d.

We followed the enantiomeric excess of **3f** formed as a function of time using CuCl, CuI and CuOTf. When using CuOTf, the reaction goes to completion much faster (<30 min) than with copper halides (overnight) and we did not detect any change in e.e. with time, so we tentatively speculate that CuOTf and Cu-halide catalysed reactions (Table 1, entry 6) are mechanistically distinct. Reactions using CuCl show a decrease of e.e. with increasing time (~82% e.e., 10 min; 68% e.e., 2 h; 54% e.e. at completion) and we speculate that when CuCl is used, the rapidly reacting enantiomer of **1d** is consumed, and not replenished by fast racemization. The reaction with the minor enantiomer therefore becomes significant and reduces the e.e. of **3f** as time progresses. We observed non-racemic **1d** (at least 11% e.e. at 2 h) during the reaction as judged by gas chromatography analysis (see Supplementary Information). In the case of CuI the initial reaction rate is slow, and only very small amounts of **3f** are formed after 10 and 30 min. Here the e.e. increases with time (~85% e.e., 10 min; 90% e.e., 2 h; 95% e.e. at completion) suggesting that a more enantioselective system is generated as the reaction progresses.

At least part of the role of CuI is to racemize **1d** via **10**, but the identity of the proposed highly enantioselective catalyst is unclear. Both CuCl and CuI give lower initial values of e.e. than are observed at later stages of the CuI-catalysed reaction. It may be that copper aggregates containing both 'Cl' and 'I' are more selective than those containing only one halogen. An experiment performed using 10% **C** + 5% CuCl + 5% CuI ambiguously gave **3f** with 90% e.e. (rather than 95% e.e.) at completion. Alternatively, salt effects may increase the e.e. of **3f** as time progresses by favouring the formation of more highly enantioselective copper-ligand complexes; the reaction produces  $\text{Cp}_2\text{ZrCl}_2$  which may affect the relative solubility of other reaction components.

<sup>1</sup>H NMR spectroscopy studies suggest that copper-ligand aggregates change during the reaction: we observe a shift in the benzylic proton signal of ligand **C**, from being characteristic of  $[\text{L}_n\text{Cu}]_n$  complexes, towards new species as time progresses (Fig. 4e, see also Supplementary Information). Cu-halide phosphoramidite aggregates exhibit complex dynamic equilibria between several solution- and solid-phase species, which are sensitive to solvent as well as halide and salt effects<sup>30</sup>, and so determining the actual composition of the copper species generated in the reaction will require further work. How the Cu-ligand complex selects for one enantiomer of **1d**, and how the Cu-ligand complex interacts with the alkylzirconium species, are also both currently unclear.

We have described a copper-catalysed enantioselective addition of alkyl zirconium reagents to racemic cyclic allylic chlorides. The reaction uses readily available starting materials and catalysts, tolerates a variety of functional groups and operates under convenient conditions. The reactions can be performed on gram scales, and we have applied the reaction to the asymmetric synthesis of biologically active cyclopentene natural products. Mechanistic studies suggest that this dynamic kinetic asymmetric transformation operates via a rapidly interconverting intermediate, racemizing the substrate, and the formation of a more highly selective copper-catalyst *in situ*. This reaction is expected to complement the well-established palladium and iridium-catalysed methods with stabilized nucleophiles. Additionally, we anticipate that the mechanistic insight will inspire future studies in the field.

Received 27 June; accepted 13 November 2014.

- Jacobsen, E. N., Pfaltz, A. & Yamamoto, H. (eds) *Comprehensive Asymmetric Catalysis: Suppl. 2* (Springer, 2004).
- Huerta, F. F., Minidis, A. B. E. & Bäckvall, J. E. Racemisation in asymmetric synthesis. Dynamic kinetic resolution and related processes in enzyme and metal catalysis. *Chem. Soc. Rev.* **30**, 321–331 (2001).
- Vedejs, E. & Jure, M. Efficiency in nonenzymatic kinetic resolution. *Angew. Chem. Int. Edn* **44**, 3974–4001 (2005).
- Trost, B. M. & VanVranken, D. L. Asymmetric transition metal-catalyzed allylic alkylations. *Chem. Rev.* **96**, 395–422 (1996).
- Trost, B. M. & Fandrick, D. R. Palladium-catalyzed dynamic kinetic asymmetric allylic alkylation with the DPPBA ligands. *Aldrichim. Acta* **40**, 59–72 (2007).

- Pfaltz, A. & Lautens, M. in *Comprehensive Asymmetric Catalysis ii* Vol. 2 (eds Jacobsen, E. N., Pfaltz, A. & Yamamoto, H.) Ch. 24, 833–884 (Springer, 1999).
- Trost, B. M. & Thaisrivongs, D. A. Strategy for employing unstabilized nucleophiles in palladium-catalyzed asymmetric allylic alkylations. *J. Am. Chem. Soc.* **130**, 14092–14093 (2008).
- Sha, S. C., Zhang, J. D., Carroll, P. J. & Walsh, P. J. Raising the  $pK_a$  limit of “soft” nucleophiles in palladium-catalyzed allylic substitutions: application of diarylmethane pronucleophiles. *J. Am. Chem. Soc.* **135**, 17602–17609 (2013).
- Lu, Z. & Ma, S. Metal-catalyzed enantioselective allylation in asymmetric synthesis. *Angew. Chem. Int. Edn* **47**, 258–297 (2008).
- Geurts, K., Fletcher, S. P., van Zijl, A. W., Minnaard, A. J. & Feringa, B. L. Copper-catalyzed asymmetric allylic substitution reactions with organozinc and Grignard reagents. *Pure Appl. Chem.* **80**, 1025–1037 (2008).
- Teichert, J. F. & Feringa, B. L. Phosphoramidites: privileged ligands in asymmetric catalysis. *Angew. Chem. Int. Edn* **49**, 2486–2528 (2010).
- Trost, B. M. & Bunt, R. C. Asymmetric induction in allylic alkylations of 3-(acyloxy)cycloalkenes. *J. Am. Chem. Soc.* **116**, 4089–4090 (1994).
- Misale, A., Niyomchon, S., Luparia, M. & Maulide, N. Asymmetric palladium-catalyzed allylic alkylation using dialkylzinc reagents: a remarkable ligand effect. *Angew. Chem. Int. Edn* **53**, 7068–7073 (2014).
- Trost, B. M. & Verhoeven, T. R. Allylic substitutions with retention of stereochemistry. *J. Org. Chem.* **41**, 3215–3216 (1976).
- Matsumita, H. & Negishi, E. Anti-stereospecificity in the palladium-catalyzed reactions of alkenyl-metal or aryl-metal derivatives with allylic electrophiles. *Chem. Commun.* 160–161 (1982).
- Harutyunyan, S. R., den Hartog, T., Geurts, K., Minnaard, A. J. & Feringa, B. L. Catalytic asymmetric conjugate addition and allylic alkylation with Grignard reagents. *Chem. Rev.* **108**, 2824–2852 (2008).
- Alexakis, A., Bäckvall, J. E., Krause, N., Pamiès, O. & Dieguez, M. Enantioselective copper-catalyzed conjugate addition and allylic substitution reactions. *Chem. Rev.* **108**, 2796–2823 (2008).
- Langlois, J. B. & Alexakis, A. in *Topics in Organometallic Chemistry* Vol. 38, *Transition Metal Catalyzed Enantioselective Allylic Substitution in Organic Synthesis* (ed. Kazmaier, U.) 235–268 (Springer, 2012).
- Norinder, J. & Bäckvall, J. E. Dynamic processes in the copper-catalyzed substitution of chiral allylic acetates leading to loss of chiral information. *Chem. Eur. J.* **13**, 4094–4102 (2007).
- Langlois, J. B. & Alexakis, A. Dynamic kinetic asymmetric transformation in copper catalyzed allylic alkylation. *Chem. Commun.* 3868–3870 (2009).
- Langlois, J. B., Emery, D., Mareda, J. & Alexakis, A. Mechanistic identification and improvement of a direct enantioconvergent transformation in copper-catalyzed asymmetric allylic alkylation. *Chem. Sci.* **3**, 1062–1069 (2012).
- Giacomina, F. & Alexakis, A. Construction of enantioenriched cyclic compounds by asymmetric allylic alkylation and ring-closing metathesis. *Eur. J. Org. Chem.* **2013**, 6710–6721 (2013).
- Maksymowicz, R. M., Roth, P. M. C. & Fletcher, S. P. Catalytic asymmetric carbon-carbon bond formation using alkenes as alkylmetal equivalents. *Nature Chem.* **4**, 649–654 (2012).
- Sidera, M., Roth, P. M. C., Maksymowicz, R. M. & Fletcher, S. P. Formation of quaternary centers by copper-catalyzed asymmetric conjugate addition of alkylzirconium reagents. *Angew. Chem. Int. Edn* **52**, 7995–7999 (2013).
- Seemann, M., Schöller, M., Kudis, S. & Helmchen, G. Syntheses of enantiomerically pure cyclopent-2-ene-1-carboxylic acid and (cyclopent-2-enyl)acetic acid by enantioselective palladium-catalyzed allylic alkylations — synthesis of enantiomerically pure (–)-chaulmoogric acid. *Eur. J. Org. Chem.* 2122–2127 (2003).
- Jacobsen, P. L. & Levy, L. Mechanism by which hydnocarpic acid inhibits mycobacterial multiplication. *Antimicrob. Agents Chemother.* **3**, 373–379 (1973).
- Cabot, M. C. & Goucher, C. R. Chaulmoogric acid-assimilation into the complex lipids of mycobacteria. *Lipids* **16**, 146–148 (1981).
- Wang, J. F. et al. Antituberculosis agents and an inhibitor of the para-aminobenzoic acid biosynthetic pathway from *Hydnocarpus anthelmintica* seeds. *Chem. Biodivers.* **7**, 2046–2053 (2010).
- Streitwieser, A., Jayasree, E. G., Hasanayn, F. & Leung, S. S. H. A theoretical study of  $\text{S}_{\text{N}}2'$  reactions of allylic halides: role of ion pairs. *J. Org. Chem.* **73**, 9426–9434 (2008).
- Zhang, H. & Gschwind, R. M. Structure identification of precatalytic copper phosphoramidite complexes in solution. *Angew. Chem. Int. Edn* **45**, 6391–6394 (2006).

Supplementary Information is available in the online version of the paper.

**Acknowledgements** We acknowledge financial support from the EPSRC (EP/H003711/1, a Career Acceleration Fellowship to S.P.F.), B. Odell and T. Claridge are thanked for assistance with the NMR experiments.

**Author Contributions** H.Y., E.R. and M.S. performed the experiments. All authors contributed to designing, analysing and discussing the experiments; S.P.F. conceived the work and guided the research. S.P.F. wrote the manuscript with assistance from H.Y. All authors contributed to discussing and editing the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.P.F. ([stephen.fletcher@chem.ox.ac.uk](mailto:stephen.fletcher@chem.ox.ac.uk)).

# The terrestrial uranium isotope cycle

Morten B. Andersen<sup>1,2</sup>, Tim Elliott<sup>1</sup>, Heye Freymuth<sup>1</sup>, Kenneth W. W. Sims<sup>3</sup>, Yaoling Niu<sup>4</sup> & Katherine A. Kelley<sup>5</sup>

Changing conditions on the Earth's surface can have a remarkable influence on the composition of its overwhelmingly more massive interior. The global distribution of uranium is a notable example. In early Earth history, the continental crust was enriched in uranium. Yet after the initial rise in atmospheric oxygen, about 2.4 billion years ago, the aqueous mobility of oxidized uranium resulted in its significant transport to the oceans and, ultimately, by means of subduction, back to the mantle<sup>1–8</sup>. Here we explore the isotopic characteristics of this global uranium cycle. We show that the subducted flux of uranium is isotopically distinct, with high  $^{238}\text{U}/^{235}\text{U}$  ratios, as a result of alteration processes at the bottom of an oxic ocean. We also find that mid-ocean-ridge basalts (MORBs) have  $^{238}\text{U}/^{235}\text{U}$  ratios higher than does the bulk Earth, confirming the widespread pollution of the upper mantle with this recycled uranium. Although many ocean island basalts (OIBs) are argued to contain a recycled component<sup>9</sup>, their uranium isotopic compositions do not differ from those of the bulk Earth. Because subducted uranium was probably isotopically unfractionated before full oceanic oxidation, about 600 million years ago, this observation reflects the greater antiquity of OIB sources. Elemental and isotope systematics of uranium in OIBs are strikingly consistent with previous OIB lead model ages<sup>10</sup>, indicating that these mantle reservoirs formed between 2.4 and 1.8 billion years ago. In contrast, the uranium isotopic composition of MORB requires the convective stirring of recycled uranium throughout the upper mantle within the past 600 million years.

Recycling of U from the surface to the Earth's deep interior can be monitored by a decrease in the Th/U ratio of the mantle<sup>2–8</sup>. Thorium provides a valuable reference for several reasons. First, U and Th behave similarly as tetravalent species in the mantle, such that they are difficult to fractionate significantly by melting processes. Only under more oxidized surface conditions do the elements show contrasting behaviour, with Th remaining tetravalent and immobile during weathering, unlike highly water-soluble hexavalent U species. Second, both Th and U are refractory, lithophile elements, and so the Th/U of the silicate Earth can be estimated from measurements of meteorites. This planetary Th/U reference has recently been refined<sup>11</sup> to a value of 3.876. The Th/U of the terrestrial upper mantle, as inferred from analyses of MORBs, is notably lower than this value; two global studies of MORB yield a mean Th/U of  $\sim 3.1$  (refs 12, 13). The low Th/U of the upper mantle is explained by addition of significant recycled U from the surface and can be reconciled with a surprisingly high time-integrated Th/U (ref. 14), as gauged from  $^{208}\text{Pb}/^{206}\text{Pb}$  ratios, if the U recycling commenced in the latter half of Earth history<sup>2–8</sup>. This makes good geological sense, as before the Great Oxidation Event (GOE)  $\sim 2.4$  Gyr ago (see, for example, ref. 15) a reduced atmosphere inhibited the surface mobility of U and prevented U recycling.

Here we test and extend this model of global U cycling using isotopic measurements of U to complement the inferences from elemental Th/U. Recent work<sup>16,17</sup> has shown that surface processes induce U isotopic variations ( $\sim 1\%$ ) that are significantly greater than the typical analytical precision ( $\sim 0.05\%$ ). Natural variations in  $^{238}\text{U}/^{235}\text{U}$  are chiefly linked to the reduction of U(VI) to U(IV), and the magnitudes of such

fractionations are inversely proportional to temperature<sup>18,19</sup>. So whereas U isotopic ratios can be perturbed at the surface, the high temperatures and dominance of tetravalent U in the mantle inhibit significant isotopic fractionations at depth. Any 'exotic'  $^{238}\text{U}/^{235}\text{U}$  signatures, produced by low-temperature fractionation and transported into the mantle, should therefore provide a robust tracer of surface-processed U.

To explore this possibility, we have characterized, to high precision, the  $\delta^{238}\text{U}$  (the parts per thousand difference in  $^{238}\text{U}/^{235}\text{U}$  relative to a reference solution standard, CRM 145) of a range of samples including meteorites, mantle-derived basalts, and the inputs and outputs of an archetypal subduction zone. A summary of our results is plotted in Fig. 1 and reported in Table 1. Measurement precision varied with sample size, but typically the magnitude of the error in  $\delta^{238}\text{U}$  was less than  $\pm 0.03\%$  (2 s.e.). For most samples, the resolution of natural variability was limited by our long-term reproducibility. This was gauged from repeat measurements of the geological standard BHVO-2, which yielded  $\delta^{238}\text{U} = -0.314\% \pm 0.028\%$  (2 s.d. on 21 replicates). We additionally obtained  $^{234}\text{U}/^{238}\text{U}$  measurements, which provide a valuable monitor of recent U disturbance. All key samples gave values within error ( $\pm 3\%$ ) of secular equilibrium for ( $^{234}\text{U}/^{238}\text{U}$ ), the  $^{234}\text{U}/^{238}\text{U}$  activity ratio. Further details of our analytical procedures and values for a wider range of standards measured are provided in Methods.

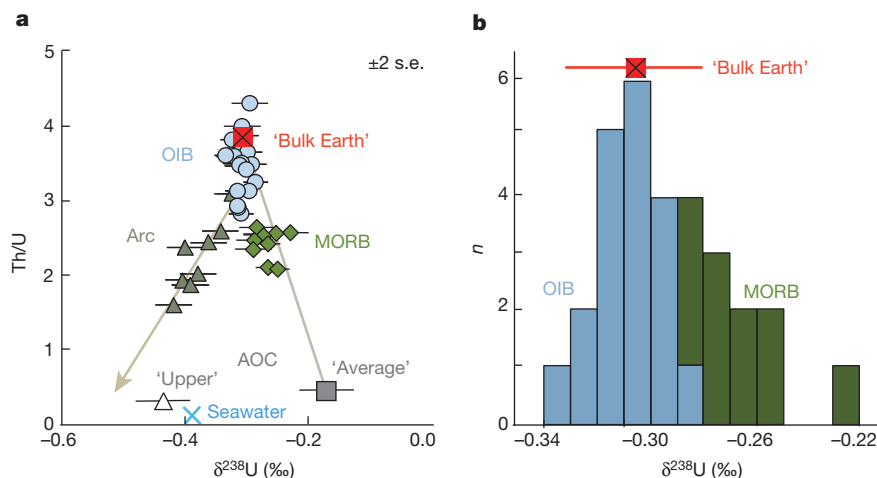
Primitive, chondritic meteorites are typically used as a reference for bulk planetary compositions, and so we analysed several ordinary chondrites to try to define a reference 'bulk Earth'  $\delta^{238}\text{U}$ . The very low U contents of chondrites make high-precision measurements especially challenging, so we further analysed two eucrites. These higher-[U] samples should still provide a useful planetary datum given the incompatible, lithophile and refractory nature of U. Our meteorite analyses have  $\delta^{238}\text{U}$  values that range from  $-0.44\%$  to  $-0.30\%$ , overlapping with two existing, lower precision determinations of chondritic  $\delta^{238}\text{U}$  ( $-0.42 \pm 0.09\%$  (ref. 20) and  $-0.37 \pm 0.09\%$  (ref. 21)). The variability in  $\delta^{238}\text{U}$  of our meteorite analyses, however, is greater than the measurement precision and probably reflects recent disturbance of some of our samples (Methods). Thus, we propose a planetary estimate based on the weighted average of the two unaltered samples, with ( $^{234}\text{U}/^{238}\text{U}$ ) within error of unity. This yields  $\delta^{238}\text{U} = -0.306 \pm 0.026\%$ .

We wish to compare the  $\delta^{238}\text{U}$  of the terrestrial mantle with this new meteoritic datum, inferred to represent bulk Earth. To this end, we have analysed a wide range of basalts (Methods), which effectively sample the mantle for an incompatible element such as U. Whereas the shallow convecting mantle is probed by MORBs, OIBs are widely assumed to be generated from hot, upwelling plumes (possibly containing components from recycled plates) that provide a window into the deeper mantle. We have analysed fresh MORB glass from all three major oceanic basins and OIBs that cover a large portion of mantle heterogeneity as gauged from radiogenic isotopic compositions. It is clear from Fig. 1 and Table 1 that MORBs and OIBs have different mean  $\delta^{238}\text{U}$ . OIBs with a wide range of Th/U have  $\delta^{238}\text{U}$  within the error of the bulk Earth, whereas MORBs have significantly higher  $\delta^{238}\text{U}$  at lower Th/U. The strikingly superchondritic  $\delta^{238}\text{U}$  we observe in MORBs would strongly support the scenario of widespread pollution of the upper mantle with

<sup>1</sup>Bristol Isotope Group, School of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK. <sup>2</sup>Institute of Geochemistry and Petrology, Department of Earth Sciences, ETH Zürich, 8092 Zürich, Switzerland.

<sup>3</sup>Department of Geology and Geophysics, University of Wyoming, Laramie, Wyoming 82071-2000, USA. <sup>4</sup>Department of Earth Sciences, Durham University, Durham DH1 3LE, UK. <sup>5</sup>Graduate School of Oceanography, University of Rhode Island, Narragansett, Rhode Island 02882-1197, USA.





**Figure 1 | Uranium isotopic compositions ( $\delta^{238}\text{U}$ ) versus Th/U ratios for mantle-derived basalts and altered oceanic crust. **a**, OIBs (circles) have  $\delta^{238}\text{U}$  similar to the bulk Earth (crossed square), whereas the higher  $\delta^{238}\text{U}$  and lower Th/U of MORBs (diamonds) imply a mixture (shown as a grey line) between bulk Earth and average modern altered oceanic crust (AOC; square, showing the ODP Site 801 supercomposite). Mariana arc basalts (triangles) show positive covariation of Th/U and  $\delta^{238}\text{U}$ , reflecting the mixing of two**

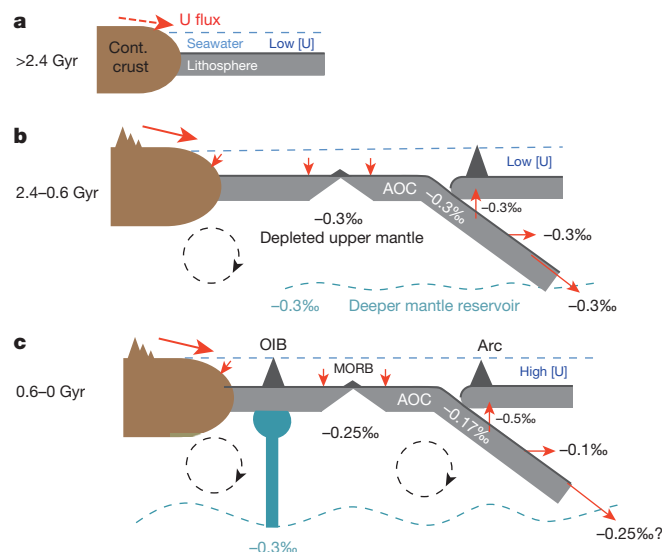
surface U, if recycled U were isotopically heavy. Thus, we examine the U isotopic composition of the subducting plate.

Although variations in  $\delta^{238}\text{U}$  in the sedimentary environment are now established, appropriate measurements to characterize subducted materials are unavailable. We principally focus on determining the isotopic composition of U added by submarine alteration to the igneous oceanic crust, which is the key flux in accounting for the low Th/U of MORBs<sup>6,8</sup>. Comprehensive studies of cores obtained from deep drilling through oceanic crust<sup>8,22,23</sup> (see also the review in ref. 24) demonstrate that heterogeneous addition of U has occurred throughout the upper ~500 m of the mafic AOC. Subduction of this 'excess' U is sufficient to lower the Th/U of the upper mantle to 2.5 in ~2 Gyr (ref. 6).

We have analysed 'composite' samples from ODP Site 801, which penetrates 420 m into ~170 Myr-old Pacific mafic crust<sup>25</sup>. The composites are mixtures of the different lithologies and alteration styles present, blended as powders in representative proportions. This is an efficient means of obtaining a reliable average composition of the heterogeneously altered mafic crust<sup>22,25</sup>. The  $\delta^{238}\text{U}$  of the composites varies from -0.45‰ in the uppermost part (0–110 m) to higher values (-0.15‰ to +0.16‰) in the deeper parts (110–420 m) (Table 1). This variability probably reflects a change from oxidic incorporation of U near the surface<sup>26</sup> to (partial) reductive U roll-front-type sequestration<sup>27</sup>, in keeping with a generally observed change in alteration style<sup>8,25</sup>. The 'supercomposite' from ODP Site 801, representing a weighted average of the full 420 m upper crustal section, has  $\delta^{238}\text{U} = -0.17‰$  and  $[\text{U}] = 0.39 \mu\text{g g}^{-1}$ ,

components: a sedimentary source, with high Th/U and  $\delta^{238}\text{U}$ ; and a source with low Th/U and low  $\delta^{238}\text{U}$  similar to the upper section of the AOC (open triangle, showing the 0–110 m ODP Site 801 composite). The brown arrow shows the best fit of the Mariana arc data pointing towards the inferred fluid component from the AOC. **b**, Histogram highlighting the distinctively heavy U isotope composition of MORBs relative to the bulk Earth and most OIBs.

which is five times higher than the unaltered basalts<sup>8</sup>. This superchondritic value for the average  $\delta^{238}\text{U}$  reflects the dominance of reductive addition of U(IV) to the upper oceanic crust as a whole (Methods). Interestingly, MORB compositions lie close to a simplistic mixing line between this average altered crustal composition and the bulk Earth reference (Fig. 1).

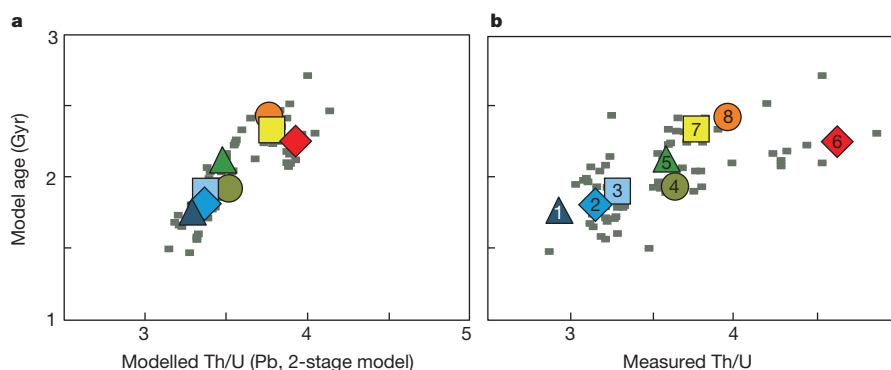


**Figure 2 | Cartoon of the terrestrial U isotope cycle over the history of Earth. **a****, Prior to the GOE ~2.4 Gyr ago, low atmospheric oxygen levels limited U mobility on the surface. **b**, The GOE would have heralded an enhanced U weathering flux to the oceans. Quantitative release of U during weathering would yield a U flux to the oceans with  $\delta^{238}\text{U} \approx -0.3‰$ , and quantitative extraction of U into the dominant reducing sinks from the ocean, including AOC, would also give  $\delta^{238}\text{U} \approx -0.3‰$  for any recycled U delivered to the mantle. **c**, During the past 600 Myr U has been isotopically fractionated during partial, reductive uptake of U into the AOC from the now largely oxic oceans. This fractionated U in the subducted AOC U flux is released at different depths; initially the isotopically light U in the uppermost crust is lost to arc magmatism, and the heavy U from the deeper crust is released beyond the arc front into the convecting upper mantle. The residual, deep-subducted crust has a U isotopic composition similar to unaltered MORB.

**Table 1 | Summary of  $\delta^{238}\text{U}$  for sample groups**

Sample type	$\delta^{238}\text{U} \pm 2 \text{ s.e.}$	Th/U	N
Unaltered meteorites (bulk Earth)	$-0.306 \pm 0.026$	3.84	2
OIBs	$-0.308 \pm 0.005$	3.48	19
MORBs	$-0.268 \pm 0.011$	2.44	11
Mafic AOC			
Average	$-0.170 \pm 0.026$	0.44	1
0–110 m	$-0.436 \pm 0.042$	0.30	3
110–220 m	$+0.164 \pm 0.086$	0.45	3
220–420 m	$-0.145 \pm 0.045$	0.44	3
Mariana arc lavas			
Guguan (fluid rich)	$-0.405 \pm 0.023$	1.73	2
Uracas (sediment rich)	$-0.333 \pm 0.016$	2.84	2
Average Mariana sediments	$-0.354 \pm 0.039$	—	7
Seawater (open ocean)	$-0.390 \pm 0.006$	—	3

A full table with individual data is presented in Supplementary Table 1.



**Figure 3 | Pb model ages versus Th/U in OIB mantle sources.** Two-stage Pb model ages ( $^{207}\text{Pb}$ – $^{206}\text{Pb}$  ages<sup>10</sup>) versus time-integrated Th/U, using measured  $^{208}\text{Pb}$ – $^{206}\text{Pb}$  and model source ages (a) and measured Th/U for the same suite of OIB samples (b): Hawaii (1), Iceland (2), Azores I (3), La Palma (4), French Polynesia (5), Samoa (6), Azores II (7) and Réunion (8)

Finally, we assess the consequences of subduction for the U isotope composition of recycled crust, using the Mariana arc as a well-understood example. During subduction, material is lost from the downgoing plate and is incorporated into magmas erupted at island arcs. The Mariana arc lavas show evidence for two slab-derived components; a melt from the sedimentary section and a ‘fluid’ from the mafic oceanic crust<sup>28</sup>. A regression line through the array of Marianas lavas in Fig. 1 should point towards the sediment component at high Th/U and the fluid component at low Th/U. The average  $\delta^{238}\text{U}$  of sediments subducting beneath the Mariana ( $-0.35\text{‰} \pm 0.04\text{‰}$ ; Table 1) is compatible with the compositions of the lavas with high Th/U, and an extension of the array of Mariana arc lavas to low Th/U indicates that the fluid derived from the mafic oceanic crust has a low  $\delta^{238}\text{U}$ . Such an isotopically light composition can be explained if either the U is preferentially lost from the uppermost altered mafic crust (Fig. 1), or U isotopes are fractionated during (partial) loss from the plate. In either case, the U that is lost to the arc is isotopically lighter than the bulk input to the subduction zone, and so U transported beyond the arc must become even heavier than its initial composition.

Thus, we are confident that modern surface cycling of U results in a substantial flux of isotopically heavy U into the mantle. This observation provides a ready explanation for the superchondritic  $\delta^{238}\text{U}$  of MORBs. The isotopically heavy U must be carried in the altered, mafic crust beyond the zone of arc magmatism, but subsequently lost to the upper mantle (Fig. 2). Transport of U in an accessory mineral such as allanite<sup>29</sup> is a possible means of effecting this outcome.

Like MORBs, many OIBs have significantly subchondritic Th/U, indicative of the addition of recycled U to their sources. Yet all OIBs have  $\delta^{238}\text{U}$  within the error of the bulk Earth value, which is inconsistent with the modern U cycle. We noted above, however, that the high  $\delta^{238}\text{U}$  characteristic of average recent oceanic crust is the result of partial reduction of U-rich, oxidized seawater as it percolates through the submarine volcanic edifice. This scenario has been possible only in the last ~600 Myr, since the second rise in oxygen in the late Proterozoic eon and the establishment of fully oxic oceans (see, for example, ref. 15). Between 600 Myr ago and the initial rise of oxygen ~2.4 Gyr ago, U was mobile during surface weathering but was rapidly scavenged from the reduced oceans<sup>15</sup>. Quantitative removal of riverine U supplied to the oceans would result in a flux of U to the sea floor that was isotopically unfractionated (Fig. 2 and Methods). The implications of this conceptual model are that OIB sources formed from recycled oceanic crust between 2.4 and 0.6 Gyr ago should have increasingly subchondritic Th/U, but chondritic  $\delta^{238}\text{U}$ . Furthermore, any OIB source formed earlier than 2.4 Gyr ago should have both chondritic Th/U and chondritic  $\delta^{238}\text{U}$ .

We have further investigated this idea using Pb model ages of several OIB suites. Their Pb model ages are calculated using a two-stage model,

(Methods). Despite potential recent disturbance of the measured Th/U in OIB samples, both plots document a similar relationship of decreasing Th/U with decreasing Pb model ages. These increasingly subchondritic Th/U ratios are consistent with progressive U addition to the mantle from subduction since the GOE ~2.4 Gyr ago.

similar to that of ref. 10 (Methods). As in ref. 10, we find that OIBs have a range of source model ages. Notably, these ages correlate with the Th/U of the samples (Fig. 3), in the manner discussed above. We show two plots in Fig. 3: one with time-integrated Th/U as determined from Pb isotopes (Fig. 3a), and the other with measured Th/U (Fig. 3b). Not unsurprisingly, the data using time-integrated Th/U form a tighter array, because any recent Th/U fractionations from source composition during melting and melt migration to the surface are removed (Methods). However, both plots independently document a similar relationship, with Th/U becoming increasingly subchondritic in OIB sources younger than ~2.4 Gyr.

The remarkable implication of the ideas presented above is that the two-stage rise in atmospheric oxygen, reconstructed from observations on the Earth’s surface, is reflected in the Th–U–Pb systematics of mantle-derived basalts. The coherence of our observations with those anticipated from a first-order model of U recycling also lends credence to the significance of Pb model ages of OIB sources<sup>10</sup>. The range of OIB model ages therefore places valuable constraints on the maximum and minimum incubation times of an OIB reservoir, which may be the longest and shortest residence times of subducted slabs at the base of the mantle before they become sufficiently buoyant to return to the surface. Furthermore, our inference that isotopically heavy U has been introduced into the mantle over only the last 600 Myr yields a maximum timescale for its effective stirring into the MORB source. This value is reassuringly consistent with an estimate based on a markedly different approach using Pb isotopes<sup>30</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 July; accepted 4 November 2014.

1. Albarède, F. & Michard, A. Transfer of continental Mg, S, O and U to the mantle through hydrothermal alteration of the oceanic crust. *Chem. Geol.* **57**, 1–15 (1986).
2. Zartman, R. E. & Haines, S. M. The plumbotectonic model for Pb isotopic systematics among major terrestrial reservoirs: a case for bi-directional transport. *Geochim. Cosmochim. Acta* **52**, 1327–1339 (1988).
3. McCulloch, M. T. The role of subducted slabs in an evolving earth. *Earth Planet. Sci. Lett.* **115**, 89–100 (1993).
4. Kramers, J. D. & Tolstikhin, I. N. Two terrestrial lead isotope paradoxes, forward transport modelling, core formation and the history of the continental crust. *Chem. Geol.* **139**, 75–110 (1997).
5. Collerson, K. D. & Kamber, B. S. Evolution of the continents and the atmosphere inferred from Th–U–Nb systematics of the depleted mantle. *Science* **283**, 1519–1522 (1999).
6. Elliott, T., Zindler, A. & Bourdon, B. Exploring the kappa conundrum: the role of recycling in the lead isotope evolution of the mantle. *Earth Planet. Sci. Lett.* **169**, 129–145 (1999).
7. Zartman, R. E. & Richardson, S. H. Evidence from kimberlitic zircon for a decreasing mantle Th/U since the Archean. *Chem. Geol.* **220**, 263–283 (2005).

8. Kelley, K. A., Plank, T., Farr, L., Ludden, J. & Staudigel, H. Subduction cycling of U, Th, and Pb. *Earth Planet. Sci. Lett.* **234**, 369–383 (2005).
9. White, W. M. & Hofmann, A. W. Sr and Nd isotope geochemistry of oceanic basalts and mantle evolution. *Nature* **296**, 821–825 (1982).
10. Chase, C. G. Oceanic island Pb: two-stage histories and mantle evolution. *Earth Planet. Sci. Lett.* **52**, 277–284 (1981).
11. Blichert-Toft, J., Zanda, B., Ebel, D. S. & Albarède, F. The Solar System primordial lead. *Earth Planet. Sci. Lett.* **300**, 152–163 (2010).
12. Gale, A., Dalton, C. A., Langmuir, C. H., Su, Y. & Schilling, J. G. The mean composition of ocean ridge basalts. *Geochem. Geophys. Geosyst.* **14**, 489–518 (2013).
13. Jenner, F. E. & O'Neill, H. S. C. Analysis of 60 elements in 616 ocean floor basaltic glasses. *Geochem. Geophys. Geosyst.* **13**, 1–11 (2012).
14. Galer, S. J. G. & O'Nions, K. Residence time of thorium, uranium and lead in the mantle with implications for mantle convection. *Nature* **316**, 778–782 (1985).
15. Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
16. Stirling, C. H., Andersen, M. B., Potter, E.-K. & Halliday, A. N. Low temperature isotope fractionation of uranium. *Earth Planet. Sci. Lett.* **264**, 208–225 (2007).
17. Weyer, S. *et al.* Natural fractionation of  $^{238}\text{U}/^{235}\text{U}$ . *Geochim. Cosmochim. Acta* **72**, 345–359 (2008).
18. Fujii, Y., Nomura, M., Onitsuka, H. & Takeda, K. Anomalous isotope fractionation in uranium enrichment processes. *J. Nucl. Sci. Technol.* **26**, 1061–1064 (1989).
19. Bigeleisen, J. Temperature dependence of the isotope chemistry of the heavy elements. *Proc. Natl Acad. Sci. USA* **93**, 9393–9396 (1996).
20. Connelly, J. N. *et al.* The absolute chronology and thermal processing of solids in the solar protoplanetary disk. *Science* **338**, 651–655 (2012).
21. Goldmann, A., Brennecke, G., Noordmann, J., Weyer, S. & Wadhwa, M. The  $^{238}\text{U}/^{235}\text{U}$  of the Earth and the Solar System. *Geochim. Cosmochim. Acta* **148**, 145–158 (2015).
22. Staudigel, H., Davies, G. R., Hart, S. R., Marchant, K. M. & Smith, B. M. Large scale isotopic Sr, Nd and O isotopic anatomy of altered oceanic crust: DSDP/ODP sites 417/418. *Earth Planet. Sci. Lett.* **130**, 169–185 (1995).
23. Bach, W., Peucker-Ehrenbrink, B., Hart, S. R. & Blusztajn, J. S. Geochemistry of hydrothermally altered oceanic crust: DSDP/ODP Hole 504B: implications for seawater-crust exchange budgets and Sr- and Pb-isotopic evolution of the mantle. *Geochem. Geophys. Geosyst.* **4**, 8904 (2003).
24. Dunk, R. M., Mills, R. A. & Jenkins, W. J. A reevaluation of the oceanic uranium budget for the Holocene. *Chem. Geol.* **190**, 45–67 (2002).
25. Kelley, K. A., Plank, T., Ludden, J. & Staudigel, H. Composition of altered oceanic crust at ODP Sites 801 and 1149. *Geochem. Geophys. Geosyst.* **4**, 8910 (2003).
26. Brennecke, G. A., Wasylenko, L. E., Bargar, J. R., Weyer, S. & Anbar, A. D. Uranium isotope fractionation during adsorption to Mn-oxyhydroxides. *Environ. Sci. Technol.* **45**, 1370–1375 (2011).
27. Bopp, C. J. IV, Lundstrom, C. C., Johnson, T. M. & Glessner, J. J. Variations in  $^{238}\text{U}/^{235}\text{U}$  in uranium ore deposits: isotopic signatures of the U reduction process? *Geology* **37**, 611–614 (2009).
28. Elliott, T., Plank, T., Zindler, A., White, W. & Bourdon, B. Element transport from slab to volcanic front at the Mariana arc. *J. Geophys. Res.* **102**, 14991–15019 (1997).
29. Hermann, J. Allanite: thorium and light rare earth element carrier in subducted crust. *Chem. Geol.* **192**, 289–306 (2002).
30. Rudge, J. F. Mantle pseudo-isochrons revisited. *Earth Planet. Sci. Lett.* **249**, 494–513 (2006).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** Financial support for this research was provided by NERC grant NE/H023933/1. We thank the Natural History Museum, London, and M. Anand for providing meteorite samples. H. Staudigel and T. Plank were instrumental in producing and curating AOC composite samples. We are grateful to C. Taylor for careful picking of MORB glasses, E. Melekhova for preparing the quenched glass, D. Vance for comments and C. Coath for maintenance of the mass spectrometers.

**Author Contributions** Analytical set-up was done by M.B.A. Sample preparation and analyses were carried out by M.B.A. and H.F. MORB samples and AOC composites were provided by K.W.W.S., Y.N. and K.A.K. All authors contributed with discussions. T.E. carried out the Pb modelling. T.E. and M.B.A. prepared the manuscript.

**Author Information** Data can be found in the EarthChem portal (<http://www.iedadata.org>). The nine-digit IGNS numbers for the sample set starts with 'IEMBA'. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.B.A. ([morten.andersen@erdw.ethz.ch](mailto:morten.andersen@erdw.ethz.ch)).



## METHODS

**Sample preparation.** With two main exceptions, existing powders of samples were used for this study. Preparation techniques for these powders are documented in the associated studies. Powdered samples of bulk meteorites were prepared at the University of Bristol for this study, from chips carefully picked under the microscope to be free from alteration, fusion crust or signs of saw-blade marks. Likewise, all MORB samples were prepared afresh by picking millimetre-sized fragments of glass without signs of alteration, surface coatings or devitrification.

Given the low U concentrations of MORB, the consequences of possible secondary U addition by absorption to marine Fe–Mn oxyhydroxide coatings is a concern. Although careful hand-picking should address this problem, we further processed the picked glass, using a mildly reductive procedure based on methods for dissolving Fe–Mn coatings in marine sediments, using a mixture of 0.05 N hydroxylamine hydrochloride, 15% acetic acid and 0.03 N Na-EDTA at pH ~4 (ref. 31). Similar approaches have been used in preparing MORB glasses for U-series analysis<sup>32–34</sup>. Each sample was leached three consecutive times by adding 12 ml of the leaching solution to the bulk samples in pre-cleaned centrifuge tubes, and placing these in a vortex shaker for 24 h at room temperature (~20 °C). Samples were thoroughly rinsed in 18 MΩ cm water following each leaching step.

The elemental concentrations of each leaching solution were determined using the Element 2 ICP-MS of the Bristol Isotope Group, using already established methods<sup>35</sup>. Potential contaminant U removed in each leach fraction was monitored using the total U concentration and the ratios of U to immobile elements Th, Sc, Ti and Zr (Supplementary Table 2). For most samples, the first leach released U in higher proportions than the immobile elements; however, at the end the three leaching steps the ratios of U to the refractory elements approached the ratios measured in the total bulk (residual) MORB sample after full dissolution. This suggests that any minor, absorbed U was effectively leached away. The U removed during reductive leaching represented only 1% to 4% of the total U that remained in the bulk residue. All MORB samples have ( $^{234}\text{U}/^{238}\text{U}$ ) within a few per mil of secular equilibrium, except D18-1, which had a ( $^{234}\text{U}/^{238}\text{U}$ ) ~14% above secular equilibrium, suggesting that some U contaminant was still present after the reductive leaching. The elevated ( $^{234}\text{U}/^{238}\text{U}$ ) of seawater (1.146) suggests that this contamination is marine derived. Despite the minor residual U contamination, we have chosen to leave it in our data set, but note its value is not as robust as the other measurements (see below for further discussion).

In addition to the natural samples, a basalt powder was made into a glass and leached to assess whether the reductive leaching caused any U isotope fractionation in the residual glass. Given the small total amounts of U removed during leaching and that the final leaching solutions appeared to have removed material with bulk glass composition (Supplementary Table 2), it seemed unlikely that leaching would bias the U isotopic composition of the residue; but for completeness we tested this. A basalt powder was melted in a platinum crucible and subsequently quenched by dropping the platinum crucible into a bath of 18 MΩ cm water at room temperature. Given the low ages of the MORB glasses in this study, they have experienced little U-series  $\alpha$ -recoil damage and can therefore be directly compared with the leaching of the artificial quenched glass. Measurements of the quenched glass gave identical  $\delta^{238}\text{U}$  before and after the reductive leaching step ( $-0.23\text{‰} \pm 0.03\text{‰}$  versus  $-0.24\text{‰} \pm 0.03\text{‰}$ ) and identical ( $^{234}\text{U}/^{238}\text{U}$ ) ( $0.995 \pm 0.003$  and  $0.999 \pm 0.002$ ), showing that the leaching process does not fractionate the U isotopes of fresh glass.

**Sample dissolution, spiking and column chemistry.** Sample sizes up to ~1.5 g were dissolved in a single beaker. For larger samples (for example chondrites and MORB) several splits of the same sample were dissolved separately and then all aliquots were combined after full dissolution. Terrestrial silicates were dissolved in a mixture of concentrated HF/HNO<sub>3</sub>/HClO<sub>4</sub> acid, fluxed on a hotplate for 24 h and dried by stepwise-increasing the temperature from 120 °C to 200 °C. Samples were then fluxed twice in ~6 N HCl (~15 ml per gram of sample) on a hotplate (120 °C) and dried down in between. Samples were redissolved in 7 N HNO<sub>3</sub> and then diluted down to 2 N HNO<sub>3</sub> for TRU Resin column chemistry. Both the eucrites (~1 g) and the ordinary chondrites (~4 g in three aliquots) were initially dissolved in a similar way to larger terrestrial silicate samples. However, after the 7 N HNO<sub>3</sub> step the dissolution was incomplete and a residue of dark residue remained. These residues were isolated from the remaining dissolved sample and refluxed in a mixture of 3 ml concentrated HNO<sub>3</sub> and 1 ml concentrated HCl in a high-pressure, Anton-Paar asher (200 °C and 100 bar pressure, in silica glass vials). This step fully dissolved the residues, which were then remixed with the already dissolved sample aliquots, dried down and redissolved in 7 N HNO<sub>3</sub> before TRU Resin chemistry in 2 N HNO<sub>3</sub>.

All samples were spiked with the IRMM-3636  $^{233}\text{U}$ – $^{236}\text{U}$  double spike<sup>36</sup> (aiming for  $^{236}\text{U}/^{235}\text{U}$  of ~5), either before dissolution or before column chemistry for samples combined from separately processed aliquots.

The U was separated from all other matrices in a two-step procedure by TRU Resin and UTEVA chemistry. The TRU Resin chemistry was optimized for large

sample sizes with ~2 ml resin loaded in polypropylene Bio-Rad columns. Up to ~1 g of sample was loaded on each column. Samples were loaded and matrix eluted using 40 ml 2 N HNO<sub>3</sub>, and U was collected in 8 ml 0.3 N HF/0.1 N HCl. Samples were then dried down and fluxed in concentrated HNO<sub>3</sub>/H<sub>2</sub>O<sub>2</sub> to eliminate any organic material from the resin bleeding into the sample. For large samples that were split over several columns, these were homogenized at the end of this phase and then dried down to be redissolved in 3 N HNO<sub>3</sub> for UTEVA chemistry. The UTEVA chemistry was performed in shrink-fit Teflon columns containing ~0.6 ml of resin. Loading and matrix elution steps used 20 ml of 3 N HNO<sub>3</sub>, before elution of Th in 3 ml of 3 N HCl, and collection of the purified U in 8 ml 0.3 N HF/0.1 N HCl. Samples were then dried down, fluxed in concentrated HNO<sub>3</sub>/H<sub>2</sub>O<sub>2</sub>, dried down and redissolved in the requisite amount of 0.2 N HCl for the desired U concentration (100–300 p.p.b.) for MC-ICPMS measurements. Full U recovery (>95%) was obtained using this method with total chemistry blanks of <20 pg for all samples (negligible compared with sample sizes >20 ng).

**MC-ICPMS measurement set-up.** The U isotope measurements were conducted on a Thermo Finnigan Neptune MC-ICP-MS (serial no. 1002) of the Bristol Isotope Group, running in low mass resolution ( $M/\Delta M \approx 500$ ) and using an Aridus desolvating nebulizer introduction system. Uranium sample sizes of 40–150 ng were consumed during individual analysis. The data were collected in static mode in a similar fashion to that described in ref. 37. All cups were connected to feedback amplifiers with 10<sup>11</sup> Ω resistors, except for the  $^{238}\text{U}$  cup, which was connected to a feedback amplifier with a 10<sup>10</sup> Ω resistor to accommodate a larger ion beam. Uranium tailing and hydride formation were monitored as described in ref. 37. Both the  $^{237.05}/^{238}\text{U}$  abundance sensitivity and the hydride and high-side tailing formation at 1 a.m.u. (measured as  $^{239.05}/^{238}\text{U}$ ) were  $(2\text{--}3) \times 10^{-6}$ , and remained stable during each measurement session. The low-mass side-tailing contributions to  $^{233}\text{U}$ ,  $^{234}\text{U}$ ,  $^{235}\text{U}$  and  $^{236}\text{U}$  were estimated, and corrected for, from interpolation of a linear–log fit to mass versus tailing intensity, as used in ref. 38. Furthermore, corrections for  $^{232}\text{ThH}^+$  and 1 a.m.u. high-mass tailing on  $^{233}\text{U}$  were made to measurement, assuming similar behaviour for Th and U. These were of minimal importance, however, owing to the good separation of U from Th during the UTEVA chemistry ( $^{232}\text{Th}/^{233}\text{U} < 1$ ).

Measurements were conducted using typical ion beam intensities of ~1 nA for  $^{238}\text{U}$ , ~7 pA for  $^{235}\text{U}$ , ~40 pA for  $^{236}\text{U}$  and  $^{233}\text{U}$ , and ~0.05 pA for  $^{234}\text{U}$ , integrated over a period of  $80 \times 4$  s. Washout and on-peak blank measurements were similar to those described in ref. 37. On-peak U blank intensity never exceeded 10 p.p.m. of the total  $^{238}\text{U}$  beam of the sample and was generally <2 p.p.m.

The Neptune was equipped with a large plasma interface pump (turbo-booster) offering enhanced transmission efficiency when combined with ‘jet+X cones’, as opposed to ‘standard+X cones’. The data obtained are identical for both set-ups (Supplementary Tables 3 and 4). General U transmission efficiencies were ~1.5% for standard+X cones and ~3% for jet+X cones.

After tailing and hydride corrections, the measured  $^{233}\text{U}/^{236}\text{U}$  ratio was used for mass bias correction, using the exponential mass fractionation law<sup>39</sup>. To obtain  $^{234}\text{U}/^{238}\text{U}$  and  $^{235}\text{U}/^{238}\text{U}$  ratios, the minute  $^{238}\text{U}$ ,  $^{235}\text{U}$  and  $^{234}\text{U}$  contributions from the IRMM-3636 spike were subtracted. Based on a calibration using CRM-145 (Supplementary Table 3), the U isotope ratios used for the Bristol IRMM-3636 were  $^{236}\text{U}/^{233}\text{U} = 0.98130$ ,  $^{236}\text{U}/^{238}\text{U} = 4,259$ ,  $^{236}\text{U}/^{235}\text{U} = 21,988$  and  $^{236}\text{U}/^{234}\text{U} = 2,770$ . These ratios are identical to the certified ratios from IRMM-3636 for  $^{236}\text{U}/^{233}\text{U}$  ( $0.98130 \pm 0.00015$ ),  $^{236}\text{U}/^{238}\text{U}$  ( $4,259 \pm 7$ ) and  $^{236}\text{U}/^{235}\text{U}$  ( $21,988 \pm 36$ ), but diverge slightly for  $^{236}\text{U}/^{234}\text{U}$  ( $2,732 \pm 4$ ). Measurements of all unknown samples were bracketed individually and normalized to CRM-145 standard measurements, spiked with IRMM-3636, in a similar fashion to the unknowns.

**Measurement performance, reproducibility and accuracy.** For each measurement sequence, the mean of the absolute  $^{238}\text{U}/^{235}\text{U}$  of the CRM-145 standard used was typically within  $\pm 50$  p.p.m. of the value (137.832) reported for the NBL 112a standard<sup>38</sup>, and the  $^{234}\text{U}/^{238}\text{U}$  ratios were within  $\pm 2\%$  of published ratios for the CRM-145 standard<sup>37</sup>. The ( $^{234}\text{U}/^{238}\text{U}$ ) activity ratios were calculated using the half-lives in ref. 40. All samples in this study were normalized to the CRM-145 standard, such that any deviations relative to the absolute ratio of the CRM-145 standard, potentially related to a non-exponential component of instrumental mass fractionation, are corrected.

Given the use of mixed feedback amplifier resistors with different response times, internal error estimates may give a misleading impression of true precision (see, for example, ref. 41). Thus, to test the full external reproducibilities for individual unknown samples, a total of eight splits of BHVO-2 were individually processed (dissolution, spiking and chemistry) and measured during different analytical sessions (with different set-ups; see Supplementary Table 4). The external reproducibility of  $\delta^{238}\text{U}$  for BHVO-2 was  $\pm 28$  p.p.m. (2 s.d.) and a similar or better external reproducibility was obtained for the basalt LP 45 E (a historic basalt from La Palma), the in-house CZ-1 uraninite and three open-ocean seawater samples (Extended Data Fig. 1). Other samples measured in duplicate agreed within their 2 s.e. estimates

(Supplementary Table 4). The reproducibilities of  $^{234}\text{U}/^{238}\text{U}$  were limited by the low  $^{234}\text{U}$  intensities ( $<0.2$  pA), but were  $\pm 3\%$  or better for the standards.

Potential artefacts in the measured  $\delta^{238}\text{U}$  from different U/matrix ratios are unlikely given that ordinary chondrites (low U/matrix ratio) and OIB (high U/matrix ratio) both have lower  $\delta^{238}\text{U}$  compared with MORB (intermediate U/matrix ratio). Furthermore, consistent results are obtained when comparing the standards measured in this study and other studies using comparable normalizing standards (NBL 112a<sup>17</sup>, SRM-950a<sup>21</sup> and CRM-145<sup>16</sup>). The measured uraninite CZ-1 standard ( $\delta^{238}\text{U} = -0.053\text{‰} \pm 0.029\text{‰}$ ) is within error of the value reported in ref. 16 ( $-0.10\text{‰} \pm 0.07\text{‰}$ ). Similarly, the BHVO-2 measurements in this study ( $\delta^{238}\text{U} = -0.314\text{‰} \pm 0.028\text{‰}$ ) compare well with measurements in ref. 21 ( $-0.32\text{‰} \pm 0.07\text{‰}$ ). Finally, our open-ocean seawater measurements ( $\delta^{238}\text{U} = -0.390\text{‰} \pm 0.018\text{‰}$ ) agree well with open-ocean seawater measurements in ref. 17 ( $-0.41\text{‰} \pm 0.03\text{‰}$ ).

**Uranium in extraterrestrial material and a bulk Earth  $\delta^{238}\text{U}$ .** Previous work has shown large variability in the  $\delta^{238}\text{U}$  of extraterrestrial material<sup>20,21,42–46</sup>. Thus, even if it exists, estimating a uniform chondritic  $\delta^{238}\text{U}$  composition is challenging. Some of the observed  $\delta^{238}\text{U}$  heterogeneity has been attributed to variable addition of  $^{235}\text{U}$  from the decay of, now extinct,  $^{247}\text{Cm}$  (ref. 45). However, meteorites also show relative depletions in  $^{235}\text{U}$ , indicating that other processes may have a role<sup>20,21,42,46</sup> (for example nucleosynthetic anomalies or planetary formation processes). An additional concern is terrestrial perturbation of U, which may be indicated from the physical preservation and anomalous chemical composition. Specifically for U, Th/U departing from the recently defined meteoritic reference<sup>11</sup> and ( $^{234}\text{U}/^{238}\text{U}$ ) activity ratios out of secular equilibrium may testify to planetary body processes<sup>47</sup> or terrestrial perturbation.

Owing to the generally low U abundance in meteorites, it is necessary to obtain large chondrite samples ( $\sim 5\text{--}10$  g) to allow high-precision  $\delta^{238}\text{U}$  measurements, which can run counter to museum loan policies. Thus, we initially honed our technique on three large desert meteorite 'finds' (M2, M12 and M15), provided by M. Anand from the Open University. These had previously been studied petrographically and characterized at the University of Bristol. We had prepared several hundred grams of powder from the interiors of these ordinary chondrites. As finds, however, these samples were potentially perturbed by terrestrial weathering. We subsequently obtained large ( $\sim 25$  g) samples of two 'falls' (Zag and Saratov) from the Meteorite Market (<http://www.meteoritemarket.com/>). All these samples are ordinary chondrites, which are not only more readily available than carbonaceous chondrites, but are isotopically more similar to the Earth (see, for example, refs 48, 49).

We supplemented our ordinary chondrite measurements with analyses of two eucrites, Juvinas and Stannern, kindly provided by the Natural History Museum, London. Eucrites have higher U contents and so require smaller sample sizes ( $\sim 1$  g) for high-precision analyses. Although differentiated meteorites, the isotope ratio of a highly incompatible element such as U should be minimally affected during crust formation, and so we believe that these samples still provide a valuable planetary reference. Notably, eucrites generally have chondritic Th/U ratios<sup>50–52</sup>, indicating an absence of elemental fractionation during their formation. The eucrite samples we analysed were falls, and so are probably less prone to terrestrial weathering than the finds.

Of the ordinary chondrites, two gave identical, but relatively low,  $\delta^{238}\text{U}$  ratios (M15 ( $-0.439\text{‰} \pm 0.030\text{‰}$ ) and Saratov ( $-0.442\text{‰} \pm 0.050\text{‰}$ )), whereas the three others were all within error but were  $\sim 100$  p.p.m. higher (M2 ( $-0.322\text{‰} \pm 0.030\text{‰}$ ), M12 ( $-0.326\text{‰} \pm 0.022\text{‰}$ ) and Zag ( $-0.301\text{‰} \pm 0.050\text{‰}$ )). The eucrites also differed in their  $\delta^{238}\text{U}$  (Juvinas ( $-0.312\text{‰} \pm 0.030\text{‰}$ ) and Stannern ( $-0.369\text{‰} \pm 0.030\text{‰}$ )), but with Juvinas overlapping with the compositions of the heaviest ordinary chondrites (Extended Data Fig. 2).

The three ordinary chondrite desert finds have elevated ( $^{234}\text{U}/^{238}\text{U}$ ) ratios (1 to 12%  $^{234}\text{U}$  excess) suggesting oxidative weathering during their time at the Earth's surface, with oxidation of Fe potentially promoting mineral surfaces for U sorption, with a positive correlation between U concentration and ( $^{234}\text{U}/^{238}\text{U}$ ) ratios (Extended Data Fig. 2). Furthermore, Saratov also had  $\sim 1\%$  elevated ( $^{234}\text{U}/^{238}\text{U}$ ), whereas Stannern was  $\sim 1\%$  depleted in ( $^{234}\text{U}/^{238}\text{U}$ ) relative to secular equilibrium. Only the ordinary chondrite Zag and eucrite Juvinas were at secular equilibrium for ( $^{234}\text{U}/^{238}\text{U}$ ). The independent constraints provided by ( $^{234}\text{U}/^{238}\text{U}$ ) show that only Juvinas and Zag can be considered pristine and, notably, their  $\delta^{238}\text{U}$  values are within error of each other (Extended Data Fig. 2).

Perturbation of U in the meteorite samples is also indicated by their Th/U ratios relative to the planetary reference value<sup>11</sup> of  $3.876 \pm 0.016$ . The ordinary chondrites with the lowest  $\delta^{238}\text{U}$  have the highest [U] and lowest Th/U, further suggestive of U addition. Three samples (Juvinas, Zag and M2) have Th/U within error of the reference value, and all have  $\delta^{238}\text{U}$  values within error of each other (Extended Data Fig. 2). However, given the minor ( $^{234}\text{U}/^{238}\text{U}$ ) excess in M2, we do not include this in our best estimate of the bulk Earth value,  $\delta^{238}\text{U} = -0.306\text{‰} \pm 0.026\text{‰}$ , provided by the weighted average and weighted 2 s.e. of Juvinas and Zag. Despite

demonstrable open-system behaviour of U, the mean of all meteorite samples gives a  $\delta^{238}\text{U}$  of  $-0.36\text{‰} \pm 0.04\text{‰}$  ( $\pm 2$  s.e.m.), which is within error of the weighted estimate from pristine samples (Extended Data Fig. 2). Although our best estimate for the bulk Earth from our meteoritic samples is defined by only two samples, and would usefully be substantiated by additional measurements, we believe that the systematics of the altered samples provide important evidence to support the significance of this best estimate. Moreover, in terms of our main observations on terrestrial samples the choice is not critical; for either the mean of all the meteorites or just the pristine ones, the  $\delta^{238}\text{U}$  of MORB are heavier while the  $\delta^{238}\text{U}$  of OIBs are unresolved from these meteoritic values.

**$\delta^{238}\text{U}$  in OIBs.** A suite of 19 OIBs from Iceland, Cape Verde, Azores, Canary Islands and Hawaii were measured. Further details on these samples are provided in Supplementary Table 1. We have dominantly used historic samples, collected previously for U-series studies, which have the major advantage of being fresh. Notably, we analysed four non-historic, but still relatively young ( $\sim 1$  Myr) and ostensibly petrographically fresh samples from La Palma, Canary Islands. Two of these samples (LPF 96-39 and CS20) yielded ( $^{234}\text{U}/^{238}\text{U}$ ) out of equilibrium, warning us against using older samples from possibly more extreme weathering environments elsewhere. Nevertheless, these two samples, showing clear open-system U-series behaviour with ( $^{234}\text{U}/^{238}\text{U}$ )  $\sim 2\%$  lower than secular equilibrium, are still within error of the other La Palma samples for  $\delta^{238}\text{U}$ , and so we did not exclude these data from our averages. This also indicates that  $\delta^{238}\text{U}$  is not hugely sensitive to minor perturbations of the U budget.

In terms of traditional radiogenic isotope characterization, the islands we have studied cover high  $^3\text{He}/^4\text{He}$  (Hawaii, Iceland), HIMU (La Palma, Canaries), EMII (Sao Miguel, Azores) and FOZO (Pico, Azores, and Fogo, Cape Verde) 'flavours' of mantle signature<sup>53–55</sup>. Although La Palma is not as radiogenic in its lead isotope ratios as the classic French Polynesian and St Helena localities, the latter have suffered  $\sim 10$  Myr of tropical weathering, and so are far from ideal for characterizing primary U isotope ratios. We have not measured any representative samples from EMII-type mantle, but nevertheless cover a large compositional range of OIB.  **$\delta^{238}\text{U}$  in MORBs.** We have measured eleven glassy, axial or near-axial MORB samples from all three major ocean basins: the Indian ( $n = 1$ ), the Atlantic ( $n = 3$ ) and the Pacific ( $n = 7$ ). Further details on these samples are given in Supplementary Table 1 and associated references<sup>56–60</sup>. All picked glasses were leached to remove possible absorbed U on ferro-manganese coating, as discussed earlier. The eleven MORB glasses have Th/U ratios of 2.1 to 2.6 and all have ( $^{234}\text{U}/^{238}\text{U}$ ) within a few per mil of secular equilibrium, except the already discussed Atlantic Ocean sample D18-1. As for the OIBs with perturbed ( $^{234}\text{U}/^{238}\text{U}$ ), the  $\delta^{238}\text{U}$  does not appear strongly affected and D18-1 has  $\delta^{238}\text{U} = -0.265\text{‰} \pm 0.030\text{‰}$ , identical to the mean  $\delta^{238}\text{U}$  of all MORB; we thus did not exclude this data point from our averages. This observation is compatible with a mass balance calculation to account for the observed ( $^{234}\text{U}/^{238}\text{U}$ ) disequilibrium of D18-1 assuming the contaminant has a U isotope composition similar to seawater. Adding seawater with ( $^{234}\text{U}/^{238}\text{U}$ ) of 1.146 to MORB at secular equilibrium should result in a change in  $\delta^{238}\text{U}$  of less than  $0.02\text{‰}$  given a seawater  $\delta^{238}\text{U}$  of  $-0.39\text{‰}$ .

**$\delta^{238}\text{U}$  in island arc volcanics.** A suite of nine mafic samples from the Mariana arc front<sup>28</sup> was selected to investigate subduction zone processes (Supplementary Table 1). These well-characterized samples show variable subducted sediment input to their sources, combined with a rather constant flux of 'fluid' from the subducting, mafic oceanic crust<sup>28</sup>. In more detail, it has recently been argued that the sediment component evident in the arc lavas is dominated by the volcanoclastic horizons rather than representing an average of all lithologies, in which pelagic clay has a significant role<sup>28,61</sup>. Samples with small sediment contributions, as marked by high  $^{143}\text{Nd}/^{144}\text{Nd}$  and low Th/Nb, have low Th/U and high ( $^{238}\text{U}/^{230}\text{Th}$ ), implying a recent, slab-derived U addition to their mantle source. The systematic compositional variations of these lavas allow us to extrapolate to the possible  $\delta^{238}\text{U}$  of this slab-derived fluid, using a best-fit linear regression line through the data in the Th/U– $\delta^{238}\text{U}$  space, as shown in Fig. 1.

**$\delta^{238}\text{U}$  of subduction zone inputs.** Subducted crust can be separated into three principal, chemical components: unaltered oceanic crust, mafic AOC and sediments. Here we analysed sediments and AOC from well-characterized deep-ocean drill holes<sup>8,25,62</sup>, ODP Site 801 and ODP Site 802 in the west Pacific, to assess the U budget of subduction-related material. Not only does the former location provide the best opportunity to assess the mean composition of the old AOC ( $\sim 170$  Myr), but because these locations are in front of the Mariana arc, the composition of the overlying sediments are specifically appropriate as the endmember for the Mariana arc lavas.

Typical assemblages for the deeper ocean sediment package that are subducted include volcanoclastics, pelagic clays, cherts, carbonates and Fe–Mn crusts. The U concentrations for these materials are variable in the  $0.1\text{--}10\text{ }\mu\text{g g}^{-1}$  range. Modern seawater has a U concentration of  $\sim 3.2\text{ ng g}^{-1}$  and a homogeneous  $\delta^{238}\text{U}$  of  $-0.390\text{‰} \pm 0.010\text{‰}$  (Supplementary Table 1). Biogenic carbonate appears to

incorporate U from seawater without any significant isotope fractionation<sup>63</sup>. The measured deep-sea pelagic clays and volcanoclastics are close to the seawater  $\delta^{238}\text{U}$  and the bulk Earth value ( $-0.42\%$  to  $-0.28\%$ ; Supplementary Table 1). Furthermore, the ODP Site 801 composite sample '801SED', meant to reflect an average of the infilling material between the pillow basalts within the basement (comprising chert, hydrothermal deposit, calcite and clay minerals) has a  $\delta^{238}\text{U}$  similar to seawater. Thus, the Mariana and indeed most subducting sediment packages have an average  $\delta^{238}\text{U}$  close to the values for modern seawater and bulk Earth.

The most important source of 'U excess' in subduction zones is the AOC. The fluid-induced alteration in oceanic crust can generally be classified into high-temperature ( $>100^\circ\text{C}$ ) and low-temperature ( $<100^\circ\text{C}$ ) types.

The high-temperature alteration generally occurs close to the spreading ridge axis and at greater depth in the crust by percolation of hot hydrothermal fluids<sup>64,65</sup>. Any seawater-derived U uptake in these settings is assumed to be quantitative (see, for example, ref. 66); however, the deeper sections of the crust ( $>1,000\text{ m}$ ) affected by high-temperature hydrothermal circulation are generally little altered and have low U concentrations close to typical MORB<sup>65</sup> (for example  $0.07\text{ }\mu\text{g g}^{-1}$ ). Thus, the high-temperature alteration at greater depth does not appear to add a significant amount of U compared with the shallower, low-temperature alteration<sup>24</sup>.

The low-temperature alteration ( $<100^\circ\text{C}$ ) dominates at ridge-flanks with percolation of less intensely heated seawater<sup>65,67</sup>, and the uppermost 500 to 1,000 m of the mafic oceanic crust experiences significant U addition<sup>8,22–25,64</sup> with a mean fivefold-greater U content relative to the unaltered MORBs<sup>8</sup>. The low-temperature alteration is therefore the cause of most uptake of additional U in the subducting plate. Thus, we have focused our attention on characterizing this low-temperature alteration using average, 'composite' samples (see below).

**Altered, mafic oceanic crust.** For estimating the U concentration and  $\delta^{238}\text{U}$  budget of altered oceanic crust, we have made high-precision  $\delta^{238}\text{U}$  analyses of 'composites' from the upper  $\sim 500\text{ m}$  of altered extrusive lavas at the well-studied ODP Site 801. This represents a substantial section of the extrusive lavas erupted at a fast spreading centre<sup>68</sup>. Secondary alteration products from hydrothermal seawater flow-through suggest alteration temperatures from 10 to  $100^\circ\text{C}$ , increasing with depth<sup>62</sup>. A typical alteration sequence consists of oxic celadonite formation around alteration veins, followed by Fe-hydroxides and then reducing saponite and pyrite, in a zone moving away from the alteration veins and into the host rock<sup>62</sup>. Carbonate precipitates also occur, which may have formed intermittently through time<sup>8,62</sup>. Uranium enrichments are evident in breccia zones and in relation to redox haloes, with U concentrated at the boundary between oxidized (celadonite-rich) and reduced (saponite/pyrite-rich) zones moving away from the alteration veins, in a roll-front-redox-type U deposition pattern<sup>8,62</sup>. These redox haloes dominate the deeper part of the drilled section<sup>8,62</sup>. In ref. 8 it is estimated that about 50% of the total U excess is hosted in the secondarily formed carbonates and that the remainder is associated with the redox haloes.

From the main (tholeiitic)  $\sim 420\text{ m}$  alteration zone, three suites of composite samples from different depth ranges (0–110 m, 110–220 m, 220–420 m) have been prepared to average the composition of the heterogeneously altered sections in the crust<sup>25</sup>. The composites are physical mixtures of powders in relative proportions of their abundances throughout the particular section of core, and are intended to physically represent the bulk composition of various depth domains within the drilled sequence. For each of the three composite zones three different powder mixtures were prepared: 'FLO' composites represent the least-altered material, 'VCL' the most altered material and 'MORB' composites represent the bulk (mixtures of the FLO and VCL composites)<sup>25</sup>. Furthermore, a 'supercomposite', comprising an integration of the full upper 420 m of core, was made<sup>25</sup>. All the composite samples have low Th/U ratios (0.1–0.6), and high U concentrations of  $\sim 0.4\text{ }\mu\text{g g}^{-1}$  (ref. 25). The U concentration of the 801 supercomposite ( $0.39\text{ }\mu\text{g g}^{-1}$ ) is similar to the DSDP 417/418 supercomposite ( $0.3\text{ }\mu\text{g g}^{-1}$ ) and significantly higher than estimated unaltered MORB<sup>69</sup> ( $0.05\text{ }\mu\text{g g}^{-1}$ ).

The  $\delta^{238}\text{U}$  was measured in the three composite sections (in all three FLO, VCL and MORB powder mixtures), the supercomposite and three individual samples. The  $\delta^{238}\text{U}$  in the composite samples are variable through the  $\sim 420\text{ m}$  sequence: the upper  $\sim 110\text{ m}$  averages  $-0.436\pm 0.042\%$ ; the middle  $\sim 110\text{ m}$  are significantly heavier, averaging  $+0.164\pm 0.086\%$ ; and the lower  $\sim 200\text{ m}$  are in between, averaging  $-0.145\pm 0.045\%$  (Supplementary Table 1). The supercomposite sample yielded a  $\delta^{238}\text{U}$  of  $-0.170\pm 0.026\%$ . In addition to the composite samples we analysed three single samples from different depths: (1) a capping alkali basalt ( $-0.333\pm 0.044\%$ ) from  $\sim 26\text{ m}$  above the 'start composite depth' of the altered crust section; (2) an altered MORB ( $-0.341\pm 0.044\%$ ) in the upper composite section ( $\sim 100\text{ m}$ ); and (3) a calcitic breccia ( $-0.114\pm 0.044\%$ ) from the lowest composite zone ( $\sim 320\text{ m}$ ). The differences in  $\delta^{238}\text{U}$  between the latter two, normal, individual altered oceanic crust samples are reassuringly consistent with the composites, with the shallower sample showing significantly lower  $\delta^{238}\text{U}$  than the deeper one. The alkali basalt sample has a high U content ( $0.7\text{ p.p.m.}$ ), presumably

reflecting its primary composition, which will be much less influenced by secondary U addition than MORB. Alkaline volcanism is atypical of oceanic crust stratigraphy but is a feature of some West Pacific drill sites, believed to be part of the burst of plume volcanism in the Cretaceous<sup>25</sup>. Fittingly, this alkali basalt sample has a  $\delta^{238}\text{U}$  similar to other OIB (Supplementary Table 1).

The variable  $\delta^{238}\text{U}$ , at values distinct from seawater, shows that the seawater-derived U is not quantitatively incorporated during alteration in the ODP Site 801 AOC. During U uptake involving no redox transition, U isotopes generally appear to yield similar or slightly lower  $\delta^{238}\text{U}$  values<sup>16,17,26,63,70</sup>. However, during the U(vi) to U(iv) reduction process U isotope fractionation is governed by both the nuclear field shift and mass-dependent mechanisms<sup>71,72</sup>. These processes lead to U isotope fractionation, but in opposite directions, with the nuclear field shift dominating the total observed  $^{238}\text{U}/^{235}\text{U}$  fractionation and leading to a preference for the heavy isotope in the reduced immobile U(iv) form<sup>18,19,71,72</sup>. Such shifts towards higher  $\delta^{238}\text{U}$ , during redox-driven U uptake, have been documented in natural environments including U-enriched reducing sediments<sup>17,37,63</sup> and redox-driven roll-front U ore deposits<sup>27,73,74</sup>. With mass balance considerations in mind, it is clear that the U incorporation constitutes a partial reduction process, because a complete reduction of the available U would result in no net isotopic fractionation. This implies a process in which U is partitioned between U(vi) and U(iv) species but only the latter is fixed, and left immobile, with the former being transported away in the percolating fluid. Such a loss of isotopically light U during a partial U reduction process preferentially taking up heavy U isotopes has been shown in groundwaters associated with roll-front U ore deposits<sup>73</sup>, *in situ* bio-stimulated U reduction flow-through experiments<sup>75</sup> and the anoxic Black Sea water column<sup>76</sup>.

For the AOC at ODP Site 801, the  $\delta^{238}\text{U}$  lower than the seawater composition in the upper 100 m may be expected from a dominant oxic U uptake, through adsorption, consistent with relatively oxidized conditions and high water/rock ratios<sup>62</sup>. A change to higher  $\delta^{238}\text{U}$  in the lower part of the AOC is in accordance with general U addition through a reductive process and a U(vi)-to-U(iv) transition in the deeper part of the AOC with more restricted seawater flow-through<sup>62,64</sup>. Furthermore, the loss of isotopically light U to the upper part of the crust will mean that fluids percolating deeper may anyway have higher  $\delta^{238}\text{U}$  signatures than the seawater composition. Both the lower and the middle part of the altered mafic crust have  $\delta^{238}\text{U}$  higher than seawater, with the highest  $\delta^{238}\text{U}$  found in the middle part. This observation may be related to the basement structure at ODP Site 801, with variable permeability and, hence, through-flow of seawater<sup>62</sup> and, consequently, heterogeneous U addition throughout the AOC. From the deposition of brown oxic haloes throughout the ODP Site 801 AOC, most oxic seawater through-flow has been estimated to occur in the upper 150 m and below 300 m depth in the core<sup>62</sup>, and, consequently, the most reducing conditions are in between. This may explain why the highest  $\delta^{238}\text{U}$  is found in the middle part, as U incorporated through U reduction is more dominant in this zone, compared to U uptake from oxic adsorption in the upper and lower sections.

Strikingly, the measured  $\delta^{238}\text{U}$  and the U concentrations are very similar for the least-altered (FLO) and most-altered (VCL) material in each of the composite sections. This suggests that it is not the degree of alteration that dictates the U incorporation, but the ambient conditions. The relatively high  $\delta^{238}\text{U}$  of the supercomposite suggests that the reduced U in the deeper part of the AOC dominates the overall  $\delta^{238}\text{U}$  signature.

Assuming that such roll-front redox U uptake, as seen in the ODP Site 801 AOC, is representative of modern AOC U uptake and represents the integrated modern AOC for subduction, it delivers a high  $\delta^{238}\text{U}$  to the mantle. In more reduced conditions U is expected to be taken up more quantitatively, resulting in little net isotopic fractionation of the added U. Such a scenario may be expected to describe the alteration of mafic oceanic crust from the percolation of anoxic seawater which dominated the deeper ocean before the second rise in atmospheric oxygen  $\sim 600\text{ Myr ago}$ <sup>15,77</sup>. This scenario would suggest insignificant U isotopic fractionation for U uptake into AOC before  $\sim 600\text{ Myr ago}$ , yielding a  $\delta^{238}\text{U}$  similar to the mean  $\delta^{238}\text{U}$  composition of rivers, the major U input into the ocean. At present, the best estimate of the modern riverine  $\delta^{238}\text{U}$  flux to the ocean is  $-0.24\%$  (ref. 78), close to our bulk Earth estimate. This suggests that U is released near-congruently with little net U isotope fractionation during oxidative terrestrial weathering and riverine transport. Assuming near-congruent U release during terrestrial weathering since the GOE  $\sim 2.4\text{ Gyr ago}$ , and oxidation of the atmosphere, quantitative uptake of U into the AOC would then imply a  $\delta^{238}\text{U}$  composition near bulk Earth in the period  $\sim 2.4\text{ Gyr}$  to  $\sim 600\text{ Myr ago}$ .

**Th–U–Pb systematics of OIB.** The database. Our simple model of U recycling predicts that samples derived from increasingly young mantle sources will have increasingly subchondritic Th/U. To test this prediction we compiled data from the literature for samples, which have been analysed for both Th and U concentrations and Pb isotope compositions (Extended Data Table 1). The latter provide model age constraints, as detailed below.



To minimize the effects of analytical problems and secondary weathering processes obscuring primary signatures, we placed quite selective criteria for inclusion of samples into the data set, as follows.

(1) We included only samples with mass-spectrometric isotope dilution data on Th and U concentrations, coupled with U-series disequilibrium data. This ensures high-precision Th/U data and provides additional information on the magnitude of possible perturbation by the melting process (see below). Moreover, because samples collected for U-series data are all young, this also guards against U perturbation during weathering. As discussed above, mobility of U during weathering is otherwise a significant concern.

(2) We have selected only data from the main, shielding-building phases of islands. Later, post-erosional lavas are frequently invoked to contain a lithospheric component<sup>79,80</sup>, and thus do not reflect the deep source we seek to investigate.

(3) In cases where several data sets exist for the sample location, we select the one containing the techniques more likely to be robust. Thus, in the case of the Azores, we use the data from ref. 81 rather than ref. 82, because the MC-ICPMS Pb data of the former provide much less-scattered model ages than those obtained from TIMS measurements of the latter.

Our database (Extended Data Table 1) thus comprises analyses from Hawaii, Iceland, Canary Islands (La Palma), Azores (Pico and São Miguel), Society Islands and Samoa. We also include a composite data point for Réunion derived from separate U-series and Pb isotope studies of historic eruptions. Although these studies are dominantly on different sample suites, the well-documented, extreme isotopic homogeneity of historic Réunion magmatism<sup>83</sup> gives us the confidence to combine the mean values of Th/U and Pb isotopes. There is one island (Pitcairn) for which appropriate data exists according to our criteria, but which we have not plotted for the following practical reasons. The very unradiogenic Pb isotope ratios of this island yields negative model ages in our calculations and so cannot be plotted together with the other data. This, combined with their extremely high Th/U, suggests that additional processes are responsible for the striking characteristics of this EMI-type composition, for example the erosion of deep continental crust<sup>84</sup>. In all, our compiled OIB database covers a similarly wide range of isotopic characteristics as represented by samples analysed for  $\delta^{238}\text{U}$  (Fig. 1), and so forms a fitting complement.

In Fig. 3 we plot each individual datum from our compilation as a point, to show the range of compositions. To emphasize the contrasting mean compositions of different islands we have also averaged individual samples from a given island. A comment is required about the averaging of the Azores samples. The island of São Miguel has a marked spatial isotopic heterogeneity, with a distinct geographic (west-to-east) variation. Thus, we have added samples from the western volcanic centre (Sete Cidades) to the Pico samples to represent 'normal' Azores (I) while the other, more easterly samples are averaged to give 'enriched' Azores (II).

Pb model ages. It has long been known that a slope on the plot of  $^{206}\text{Pb}/^{204}\text{Pb}$  versus  $^{207}\text{Pb}/^{204}\text{Pb}$  potentially has age significance (see, for example, ref. 85). This approach was used to some effect in ref. 10, using the linear arrays in  $^{206}\text{Pb}/^{204}\text{Pb}$  versus  $^{207}\text{Pb}/^{204}\text{Pb}$  defined by some OIBs to calculate isochron ages of their sources, which ranged from 1 to 2.5 Gyr. Here we follow a similar approach. However, we did not want to rely on islands yielding well-defined linear arrays in Pb isotope space. Instead, we calculate the model ages of individual points rather than the slope of an array of data. Both approaches assume a common first stage for all samples. From this evolving reservoir a secondary model age ( $t_m$ ) and U/Pb ( $\mu_2$ ) are calculated to produce the modern Pb isotopic composition. The parameters of our first-stage evolution are given in Extended Data Table 2 together with other input values.

The Pb model age we calculate represents an event that increased U/Pb to generate modern Pb isotopic compositions that lie to the right of the geochron. As discussed widely (see, for example, ref. 86), the process of subduction provides an appealing physical manifestation of this model scenario. During subduction, dehydration preferentially removes Pb from the mafic crust<sup>87</sup>, increasing the U/Pb and Th/Pb of the deep subducted residue. Thus, we believe that the model ages relate to the time of subduction of recycled oceanic crust found in OIB sources.

Explicitly, we calculate our model ages by rearranging and numerically solving the following two equations (1) and (2) for  $\mu_2$  and  $t_m$  (using parameters described in Extended Data Table 2):

$$\frac{^{206}\text{Pb}}{^{204}\text{Pb}} = \frac{^{206}\text{Pb}}{^{204}\text{Pb}_{\text{CD}}} + \mu_1 (e^{\lambda_{238}t} - e^{\lambda_{238}t_m}) + \mu_2 (e^{\lambda_{238}t_m} - 1) \quad (1)$$

$$\frac{^{207}\text{Pb}}{^{204}\text{Pb}} = \frac{^{207}\text{Pb}}{^{204}\text{Pb}_{\text{CD}}} + \frac{\mu_1}{137.88} (e^{\lambda_{235}t} - e^{\lambda_{235}t_m}) + \frac{\mu_2}{137.88} (e^{\lambda_{235}t_m} - 1) \quad (2)$$

Having obtained  $\mu_2$  and  $t_m$  a model Th/U (weight ratio) of the second stage may then be calculated accordingly by rearranging equation (3):

$$\frac{^{208}\text{Pb}}{^{204}\text{Pb}} = \frac{^{208}\text{Pb}}{^{204}\text{Pb}_{\text{CD}}} + \frac{\text{Th}}{\text{U}} \kappa \mu_1 (e^{\lambda_{232}t} - e^{\lambda_{232}t_m}) + \frac{\text{Th}}{\text{U}} \kappa \mu_2 (e^{\lambda_{232}t_m} - 1) \quad (3)$$

Extended Data Table 1 contains the averaged, calculated model Pb ages for our OIB data set, whereas the individual model Pb ages versus measured Th/U and modelled Th/U (Pb, two-stage) are shown in Fig. 3. We use Th/U (weight ratio) throughout as this is most commonly reported in the literature, although  $^{232}\text{Th}/^{238}\text{U}$  (atomic ratio,  $\kappa$ ) is required in the calculations (the difference between these two ratios is not great). The measured Th/U is potentially perturbed by melting or during melt migration to the surface, or both. For this reason, we used only samples for which U-series measurements were available, which provides direct constraints on the magnitude of this process. All samples in our data set have  $^{230}\text{Th}$  excesses, from 1% to 37% (Extended Data Fig. 3). This implies no more than a 37% increase in Th/U during melt generation and likely less (see, for example, the discussion in ref. 88). As in the case of MORB, recent melting cannot explain the trend to lower Th/U from values close to the planetary reference (3.876).

The Samoan samples are notable for having Th/U higher (4.0–5.3) than the planetary reference value. This cannot be solely a result of recent melt fractionation because these samples have minor ( $^{230}\text{Th}/^{238}\text{U}$ ) disequilibrium (Extended Data Fig. 3). This high Th/U is potentially associated with lithospheric enrichment from plume-derived carbonatitic metasomatism, which can fractionate Th/U but will not influence the Pb isotopes<sup>89</sup>. We note that the Samoan Pb isotopes are incompatible with their high Th/U being associated with ancient fractionation; that is, they have model Th/U within error of the planetary Th/U. Although the Samoan source has long been associated with recycled continental sediments<sup>90</sup>, if these were recycled before the major rise in atmospheric oxygen (as is compatible with their model ages), then the Th/U of the continental material should be unfractionated<sup>91,92</sup>. Consequently, whether or not the enriched component is continental is not a critical issue. We also stress that the overall trend in Fig. 3 is not pinned by Samoa, but includes Réunion and the enriched samples of São Miguel. For the latter, there has been a detailed discussion of why recycled sediment is not implicated in this source<sup>28</sup>.

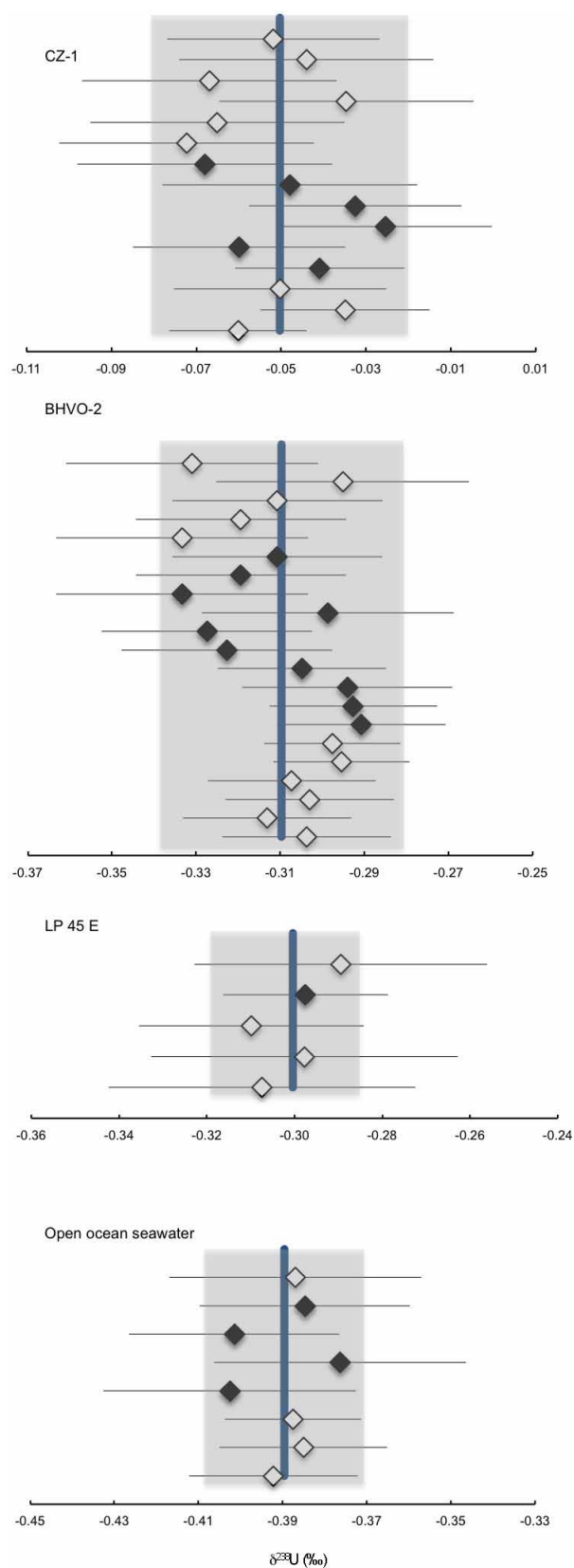
The apparently continuously declining Th/U of OIBs and their typically higher values than MORB argues against excess U left residual in the subducting slab having a significant role in lowering Th/U of OIBs. Rather, we infer that the slab adds its excess U, from sea-floor alteration, to the upper mantle. As hypothesized in the main text this is likely to result from its mineralogical host becoming unstable during pro-grade metamorphism. A host such as allanite<sup>29</sup> would survive beyond the subduction zone, but would ultimately melt to transfer U into the surrounding mantle. Thus, we infer that the declining Th/U of OIBs reflects the steadily decreasing Th/U of the upper-mantle source, which forms crust subsequently recycled to produce further OIBs. In this model, the upper mantle always has lower Th/U than previously formed OIB sources. The need for the excess U to be lost from recycled oceanic crust has also been discussed in terms of the Pb isotope systematics of OIB (see, for example, refs 93, 94).

31. Gutjahr, M. *et al.* Reliable extraction of a deepwater trace metal isotope signal from Fe–Mn oxyhydroxide coatings of marine sediments. *Chem. Geol.* **242**, 351–370 (2007).
32. Goldstein, S. J., Murrell, M. T. & Janecky, D. R. Th and U isotopic systematics of basalts from the Juan de Fuca and Gorda Ridges by mass spectrometry. *Earth Planet. Sci. Lett.* **96**, 134–146 (1989).
33. Bourdon, B., Goldstein, S. J., Bourles, D., Murrell, M. T. & Langmuir, C. H. Evidence from  $^{10}\text{Be}$  and U series disequilibria on the possible contamination of mid-ocean ridge basalt glasses by sedimentary material. *Geochim. Geophys. Geosyst.* **1**, 2000GC000047 (2000).
34. Reinartz, I. & Turekian, K. K.  $^{230}\text{Th}/^{238}\text{U}$  and  $^{226}\text{Ra}/^{230}\text{Th}$  fractionation in young basaltic glasses from the East Pacific Rise. *Earth Planet. Sci. Lett.* **94**, 199–207 (1989).
35. Andersen, M. B., Vance, D., Keech, A. R., Rickli, J. & Hudson, G. Estimating U fluxes in a high-latitude, boreal post-glacial setting using U-series isotopes in soils and rivers. *Chem. Geol.* **354**, 22–32 (2013).
36. Richter, S. *et al.* The isotopic composition of natural uranium samples—Measurements using the new  $^{233}\text{U}/^{236}\text{U}$  double spike IRMM-3636. *Int. J. Mass Spectrom.* **269**, 145–148 (2008).
37. Andersen, M. B. *et al.* A modern framework for the interpretation of  $^{238}\text{U}/^{235}\text{U}$  in studies of ancient ocean redox. *Earth Planet. Sci. Lett.* **400**, 184–194 (2014).
38. Hiess, J., Condon, D. J., McLean, N. & Noble, S. R.  $\text{U}^{238}/\text{U}^{235}$  systematics in terrestrial uranium-bearing minerals. *Science* **335**, 1610–1614 (2012).
39. Russell, W. A., Papanastassiou, D. & Tombrello, T. A. Ca isotope fractionation on the Earth and other solar system materials. *Geochim. Cosmochim. Acta* **42**, 1075–1090 (1978).
40. Cheng, H. *et al.* Improvements in  $^{230}\text{Th}$  dating,  $^{230}\text{Th}$  and  $^{234}\text{U}$  half-life values, and U–Th isotopic measurements by multi-collector inductively coupled plasma mass spectrometry. *Earth Planet. Sci. Lett.* **371–372**, 82–91 (2013).
41. Steele, R. C. J., Elliott, T., Coath, C. D. & Regelous, M. Confirmation of mass-independent Ni isotopic variability in iron meteorites. *Geochim. Cosmochim. Acta* **75**, 7906–7925 (2011).

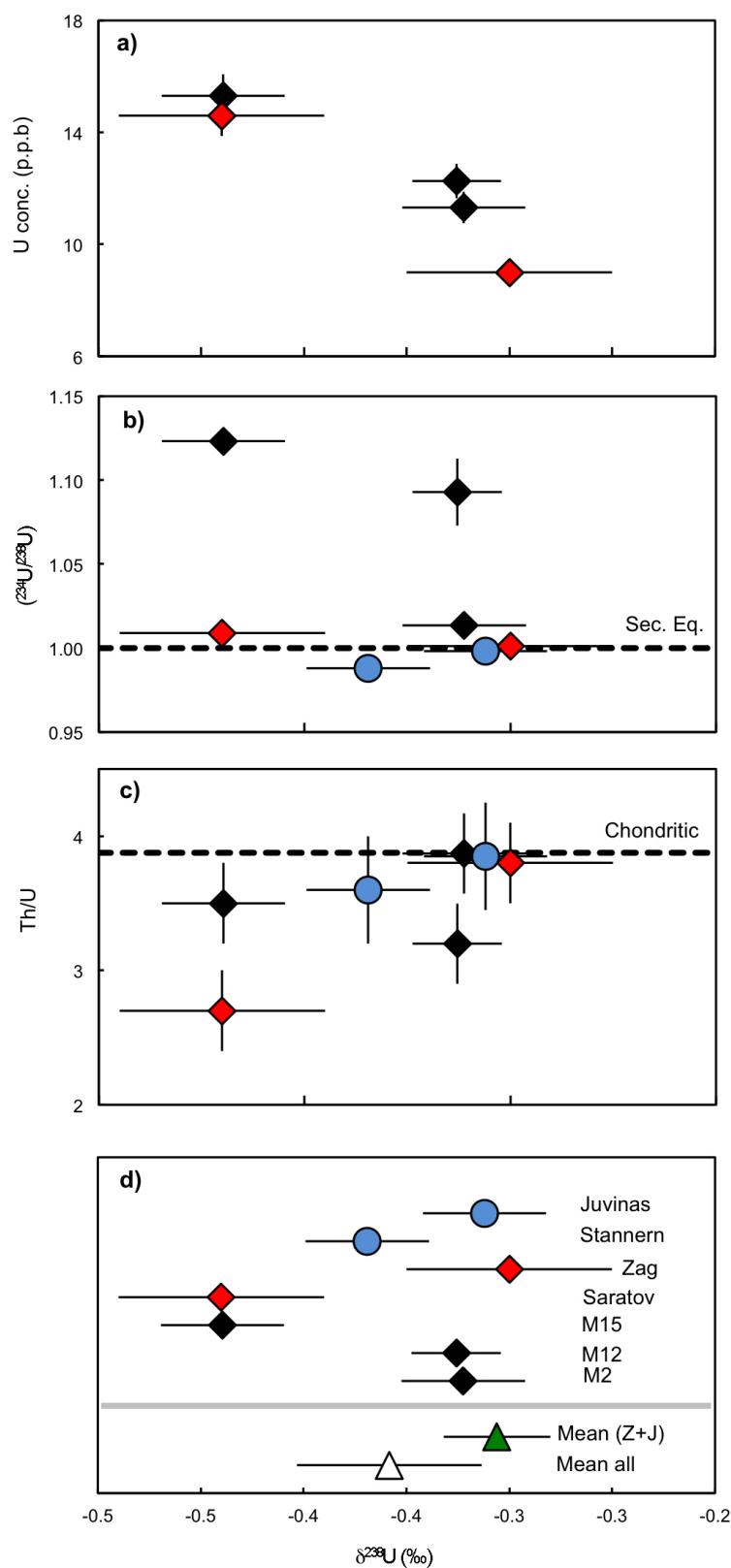
42. Stirling, C. H., Halliday, A. N. & Porcelli, D. In search of live  $^{247}\text{Cm}$  in the early solar system. *Geochim. Cosmochim. Acta* **69**, 1059–1071 (2005).
43. Stirling, C. H., Halliday, A. N., Potter, E.-K., Andersen, M. B. & Zanda, B. A low initial abundance of  $^{247}\text{Cm}$  in the early solar system: implications for r-process nucleosynthesis. *Earth Planet. Sci. Lett.* **251**, 386–397 (2006).
44. Brennecka, G. A. & Wadhwa, M. Uranium isotope compositions of the basaltic angrite meteorites and the chronological implications for the early Solar System. *Proc. Natl Acad. Sci. USA* **109**, 9299–9303 (2012).
45. Brennecka, G. A. *et al.*  $^{238}\text{U}/^{235}\text{U}$  variations in meteorites: extant  $^{247}\text{Cm}$  and implications for Pb–Pb dating. *Science* **327**, 449–451 (2010).
46. Amelin, Y. *et al.* U–Pb chronology of the Solar System's oldest solids with variable  $^{238}\text{U}/^{235}\text{U}$ . *Earth Planet. Sci. Lett.* **300**, 343–350 (2010).
47. Rocholl, A. & Jochum, K. P. Th, U and other trace-elements in carbonaceous chondrites: implications for the terrestrial and solar-system Th/U ratios. *Earth Planet. Sci. Lett.* **117**, 265–278 (1993).
48. Dauphas, N., Marty, B. & Reisberg, L. Molybdenum evidence for inherited planetary scale isotope heterogeneity of the protosolar nebula. *Astrophys. J.* **565**, 640–644 (2002).
49. Trinquier, A. *et al.* Origin of nucleosynthetic isotope heterogeneity in the solar protoplanetary disk. *Science* **324**, 374–376 (2009).
50. Barrat, J. A. *et al.* The Stannern trend eucrites: contamination of main group eucritic magmas by crustal partial melts. *Geochim. Cosmochim. Acta* **71**, 4108–4124 (2007).
51. Morgak, J. W. & Lovering, J. F. Uranium and thorium in achondrites. *Geochim. Cosmochim. Acta* **37**, 1697–1707 (1973).
52. Manhès, G., Allègre, C. J. & Provost, A. U–Th–Pb systematics of the eucrite “Juvinas”: precise age determination and evidence for exotic lead. *Geochim. Cosmochim. Acta* **48**, 2247–2264 (1984).
53. Zindler, A. & Hart, S. Chemical geodynamics. *Annu. Rev. Earth Planet. Sci.* **14**, 493–571 (1986).
54. Hart, S. R., Hauri, E. H., Oschmann, L. A. & Whitehead, J. A. Mantle plumes and entrainment: isotopic evidence. *Science* **256**, 517–520 (1992).
55. Farley, K. A. & Neroda, E. Noble gases in the Earth's mantle. *Annu. Rev. Earth Planet. Sci.* **26**, 189–218 (1998).
56. Sims, K. W. W. *et al.* Chemical and isotopic constraints on the generation and transport of magma beneath the East Pacific Rise. *Geochim. Cosmochim. Acta* **66**, 3481–3504 (2002).
57. Waters, C. L. *et al.* Recent volcanic accretion at  $9^\circ\text{N}$ – $10^\circ\text{N}$  East Pacific Rise as resolved by combined geochemical and geological observations. *Geochim. Geophys. Geosyst.* **14**, 2547–2574 (2013).
58. Regelous, M. *et al.* Variations in the geochemistry of magmatism on the East Pacific Rise at  $10^\circ 30'\text{N}$  since 800 ka. *Earth Planet. Sci. Lett.* **168**, 45–63 (1999).
59. Regelous, M., Niu, Y., Abouchami, W. & Castillo, P. R. Shallow origin for South Atlantic Dupal Anomaly from lower continental crust: geochemical evidence from the Mid-Atlantic Ridge at 26 S. *Lithos* **112**, 57–72 (2009).
60. Robinson, C. J., White, R. S., Bickle, M. J. & Minshull, T. A. Restricted melting under the very slow-spreading Southwest Indian Ridge. *Geol. Soc. Lond. Spec. Publ.* **118**, 131–141 (1996).
61. Avanzinelli, R. *et al.* Combined  $^{238}\text{U}/^{230}\text{Th}$  and  $^{235}\text{U}/^{231}\text{Pa}$  constraints on the transport of slab-derived material beneath the Mariana Islands. *Geochim. Cosmochim. Acta* **92**, 308–328 (2012).
62. Alt, J. C. & Teagle, D. A. Hydrothermal alteration of upper oceanic crust formed at a fast-spreading ridge: mineral, chemical, and isotopic evidence from ODP Site 801. *Chem. Geol.* **201**, 191–211 (2003).
63. Romaniello, S. J., Herrmann, A. D. & Anbar, A. D. Uranium concentrations and  $^{238}\text{U}/^{235}\text{U}$  isotope ratios in modern carbonates from the Bahamas: assessing a novel paleoredox proxy. *Chem. Geol.* **362**, 305–316 (2013).
64. Alt, J. C. *et al.* Subsurface structure of a submarine hydrothermal system in ocean crust formed at the East Pacific Rise, ODP/IODP Site 1256. *Geochim. Geophys. Geosyst.* **11**, 2010GC003144 (2010).
65. Staudigel, H. Hydrothermal alteration processes in the oceanic crust. *Treatise Geochem.* **3**, 511–535 (2003).
66. Chen, J., Wasserburg, G., Von Damm, K. & Edmond, J. The U–Th–Pb systematics in hot springs on the East Pacific Rise at 21 N and Guaymas Basin. *Geochim. Cosmochim. Acta* **50**, 2467–2479 (1986).
67. Mottl, M. *et al.* Warm springs discovered on 3.5 Ma oceanic crust, eastern flank of the Juan de Fuca Ridge. *Geology* **26**, 51–54 (1998).
68. Plank, T. *et al.* *Proc. Ocean Drilling Program, Initial Reports Vol. 185* (Ocean Drilling Program, 2000).
69. Staudigel, H., Plank, T., White, B. & Schmincke, H.-U. Geochemical fluxes during seafloor alteration of the basaltic upper oceanic crust: DSDP Sites 417 and 418. *Geophys. Monogr. Ser.* **96**, 19–38 (1996).
70. Shiel, A. E. *et al.* No measurable changes in  $^{238}\text{U}/^{235}\text{U}$  due to desorption–adsorption of U(VI) from groundwater at the Rifle, Colorado, integrated field research challenge site. *Environ. Sci. Technol.* **47**, 2535–2541 (2013).
71. Bigeleisen, J. Nuclear size and shape effects in chemical reactions. Isotope chemistry of heavy elements. *J. Am. Chem. Soc.* **118**, 3676–3680 (1996).
72. Fujii, Y., Higuchi, N., Haruno, Y., Nomura, M. & Suzuki, T. Temperature dependence of isotope effects in uranium chemical exchange reactions. *J. Nucl. Sci. Technol.* **43**, 400–406 (2006).
73. Murphy, M. J., Stirling, C. H., Kaltenbach, A., Turner, S. P. & Schaefer, B. F. Fractionation of  $^{238}\text{U}/^{235}\text{U}$  by reduction during low temperature uranium mineralisation processes. *Earth Planet. Sci. Lett.* **388**, 306–317 (2014).
74. Brennecka, G. A., Borg, L. E., Hutcheon, I. D., Sharp, M. A. & Anbar, A. D. Natural variations in uranium isotope ratios of uranium ore concentrates: understanding the  $^{238}\text{U}/^{235}\text{U}$  fractionation mechanism. *Earth Planet. Sci. Lett.* **291**, 228–233 (2010).
75. Bopp, C. J., IV *et al.* Uranium  $^{238}\text{U}/^{235}\text{U}$  isotope ratios as indicators of reduction: results from an in situ biostimulation experiment at Rifle, Colorado, USA. *Environ. Sci. Technol.* **44**, 5927–5933 (2010).
76. Romaniello, S. J., Brennecka, G. A., Anbar, A. D. & Colman, A. S. Natural isotopic fractionation of  $^{238}\text{U}/^{235}\text{U}$  in the water column of the Black Sea. *Eos Trans. AGU* **90**, 52, V54C–06 (2009).
77. Partin, C. A. *et al.* Large-scale fluctuations in Precambrian atmospheric and oceanic oxygen levels from the record of U in shales. *Earth Planet. Sci. Lett.* **369**, 284–293 (2013).
78. Noordmann, J. *et al.* Fractionation of  $^{238}\text{U}/^{235}\text{U}$  during weathering and hydrothermal alteration. *Mineral. Mag.* **76**, A1548 (2012).
79. Class, C. & Goldstein, S. L. Plume–lithosphere interactions in the ocean basins: constraints from the source mineralogy. *Earth Planet. Sci. Lett.* **150**, 245–260 (1997).
80. Lundstrom, C., Hoernle, K. & Gill, J. U-series disequilibria in volcanic rocks from the Canary Islands: plume versus lithospheric melting. *Geochim. Cosmochim. Acta* **67**, 4153–4177 (2003).
81. Elliott, T., Blichert-Toft, J., Heumann, A., Koetsier, G. & Forjaz, V. The origin of enriched mantle beneath Sao Miguel, Azores. *Geochim. Cosmochim. Acta* **71**, 219–240 (2007).
82. Turner, S., Hawkesworth, C., Rogers, N. & King, P. U–Th isotope disequilibria and ocean island basalt generation in the Azores. *Chem. Geol.* **139**, 145–164 (1997).
83. Graham, D., Lupton, J., Albarède, F. & Condomines, M. Extreme temporal homogeneity of helium-isotopes at Piton-De-La-Fournaise, Réunion Island. *Nature* **347**, 545–548 (1990).
84. Willbold, M. & Stracke, A. Trace element composition of mantle endmembers: implications for recycling of oceanic and upper and lower continental crust. *Geochim. Geophys. Geosyst.* **7**, 2005GC001005 (2006).
85. Patterson, C. C. Age of meteorites and the Earth. *Geochim. Cosmochim. Acta* **10**, 230–237 (1956).
86. Chauvel, C., Lewin, E., Carpentier, M., Arndt, N. T. & Marini, J.-C. Role of recycled oceanic basalt and sediment in generating the Hf–Nd mantle array. *Nature Geosci.* **1**, 64–67 (2008).
87. Miller, D. M., Goldstein, S. L. & Langmuir, C. H. Cerium/lead and lead isotope ratios in arc magmas and the enrichment of lead in the continents. *Nature* **368**, 514–520 (1994).
88. Elliott, T. Fractionation of U and Th during mantle melting: a reprise. *Chem. Geol.* **139**, 165–183 (1997).
89. Hauri, E. H., Shimizu, N., Dieu, J. J. & Hart, S. R. Evidence for hotspot-related carbonatite metasomatism in the oceanic upper mantle. *Nature* **365**, 221–227 (1993).
90. Wright, E. & White, W. M. The origin of Samoa: new evidence from Sr, Nd, and Pb isotopes. *Earth Planet. Sci. Lett.* **81**, 151–162 (1987).
91. McLennan, S. M. & Taylor, S. R. Th and U in sedimentary rocks: crustal evolution and sedimentary recycling. *Nature* **285**, 621–624 (1980).
92. Jackson, M. G. *et al.* The return of subduction continental crust in Samoan lavas. *Nature* **448**, 684–687 (2007).
93. Staudigel, H. & Hart, S. R. Alteration of basaltic glass: mechanisms and significance for the oceanic crust–seawater budget. *Geochim. Cosmochim. Acta* **47**, 337–350 (1983).
94. Chauvel, C., Hofmann, A. W. & Vidal, P. HIMU-EM: the French Polynesian connection. *Earth Planet. Sci. Lett.* **110**, 99–119 (1992).
95. Pietruszka, A. J. & Garcia, M. O. The size and shape of Kilauea Volcano's summit magma storage reservoir: a geochemical probe. *Earth Planet. Sci. Lett.* **167**, 311–320 (1999).
96. Sims, K. W. W. *et al.* Mechanisms of magma generation beneath Hawaii and mid-ocean ridges: uranium/thorium and samarium/neodymium isotopic evidence. *Science* **267**, 508–512 (1995).
97. Sims, K. W. W. *et al.* Porosity of the melting zone and variations in the solid mantle upwelling rate beneath Hawaii: inferences from  $^{238}\text{U}$ ,  $^{230}\text{Th}$ ,  $^{226}\text{Ra}$  and  $^{235}\text{U}$ ,  $^{231}\text{Pa}$  disequilibria. *Geochim. Cosmochim. Acta* **63**, 4119–4138 (1999).
98. Kokfelt, T. F. *et al.* Combined trace element and Pb–Nd–Sr–O isotope evidence for recycled oceanic crust (upper and lower) in the Iceland mantle plume. *J. Petrol.* **47**, 1705–1749 (2006).
99. Kokfelt, T. F., Hoernle, K. & Hauff, F. Upwelling and melting of the Iceland plume from radial variation of  $^{238}\text{U}$ ,  $^{230}\text{Th}$  disequilibria in postglacial volcanic rocks. *Earth Planet. Sci. Lett.* **214**, 167–186 (2003).
100. Prytulak, J. & Elliott, T. Determining melt productivity of mantle sources from  $^{238}\text{U}$ ,  $^{230}\text{Th}$  and  $^{235}\text{U}$ ,  $^{231}\text{Pa}$  disequilibria: an example from Pico Island, Azores. *Geochim. Cosmochim. Acta* **73**, 2103–2122 (2009).
101. Prytulak, J. *et al.* Melting versus contamination effects on  $^{238}\text{U}$ ,  $^{230}\text{Th}$ ,  $^{226}\text{Ra}$  and  $^{235}\text{U}$ ,  $^{231}\text{Pa}$  disequilibria in lavas from Sao Miguel, Azores. *Chem. Geol.* **381**, 94–109 (2014).
102. Elliott, T. *Element Fractionation in the Petrogenesis of Ocean Island Basalts* 29–92. PhD thesis, Open Univ. (1991).
103. Marcantonio, F., Zindler, A., Elliott, T. & Staudigel, H. Os isotope systematics of La Palma, Canary Islands: evidence for recycled crust in the mantle source of HIMU ocean islands. *Earth Planet. Sci. Lett.* **133**, 397–410 (1995).
104. Hémond, C., Devey, C. W. & Chauvel, C. Source compositions and melting processes in the Society and Austral plumes (South Pacific Ocean): element and isotope (Sr, Nd, Pb, Th) geochemistry. *Chem. Geol.* **115**, 7–45 (1994).
105. Sims, K. W. W. & Hart, S. R. Comparison of Th, Sr, Nd and Pb isotopes in oceanic basalts: implications for mantle heterogeneity and magma genesis. *Earth Planet. Sci. Lett.* **245**, 743–761 (2006).

106. Bosch, D. *et al.* Pb, Hf and Nd isotope compositions of the two Réunion volcanoes (Indian Ocean): a tale of two small-scale mantle “blobs”? *Earth Planet. Sci. Lett.* **265**, 748–765 (2008).
107. Sigmarsson, O., Condomines, M. & Bachèlery, P. Magma residence time beneath the Piton de la Fournaise Volcano, Reunion Island, from U-series disequilibria. *Earth Planet. Sci. Lett.* **234**, 223–234 (2005).





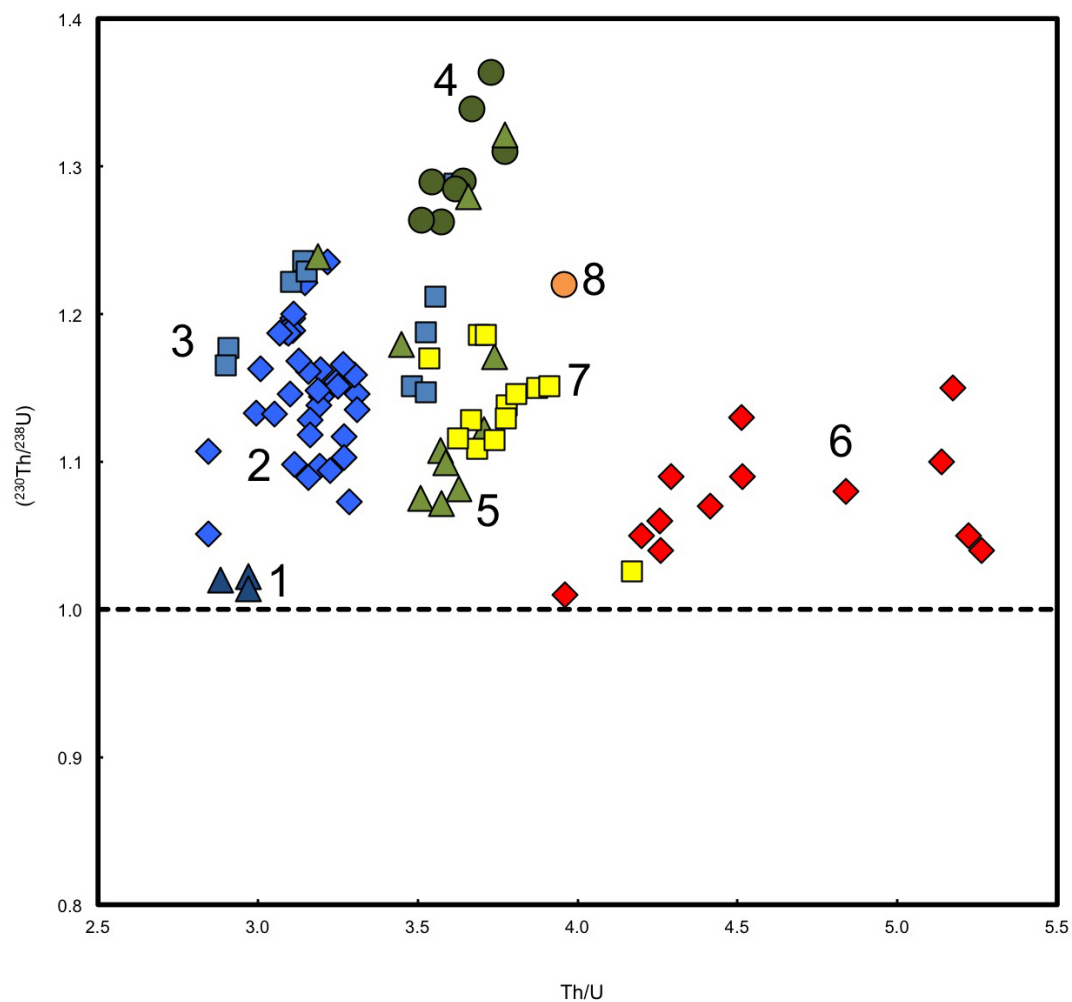
**Extended Data Figure 1 |  $\delta^{238}\text{U}$  reproducibility of standards.** Repeated  $\delta^{238}\text{U}$  measurements of a range of standards with different matrixes (CZ-1 uraninite, BHVO-2/LP 45 E basalts, seawater) are shown. All have external reproducibility (2 s.d., grey shaded area) better than  $\pm 0.30\text{‰}$ , a similar range to the internal measurement uncertainty (2 s.e.) for individual samples (Methods). The different symbols refer to the different measurement set-ups (Supplementary Table 4).



### Extended Data Figure 2 | U–Th geochemistry of analysed meteorites.

**a.**  $\delta^{238}\text{U}$  versus U concentration for ordinary chondrites (black diamonds, 'finds'; red diamonds, 'falls'). **b.**  $\delta^{238}\text{U}$  versus  $(^{234}\text{U}/^{238}\text{U})$  for ordinary chondrites (symbols as in **a**) and eucrites (blue circles). **c.**  $\delta^{238}\text{U}$  versus Th/U for

the same samples as in **a** and **b**. **d.** A 'Caltech plot' of the  $\delta^{238}\text{U}$  of individual meteorite samples and averages based on (1) the only two meteorites with  $(^{234}\text{U}/^{238}\text{U})$  within error of secular equilibrium ('Mean (Z+J)') and (2) all of the analysed meteorites ('Mean all'). Error bars denote 2 s.e.m.



**Extended Data Figure 3 | U–Th isotope systematics in the OIB used for Pb age modelling.** Symbol colours are as in Fig. 3: (1) Hawaii, (2) Iceland, (3) Azores I, (4) La Palma, (5) French Polynesia, (6) Samoa, (7) Azores II,

(8) Réunion. References can be found in Extended Data Table 1. Note that the y axis shows activity ratio whereas the x axis shows a weight ratio. The dashed line represents secular equilibrium of  $(^{230}\text{Th}/^{238}\text{U})$ .



**Extended Data Table 1 | Literature compilation of Pb, U and Th in Ocean Island Basalts**

N*	Locality	Ref. Pb	Ref. Th/U	n†	age (Ga)‡
1.	Hawaii (Kilauea)	(95,96,97)	(95,96,97)	3	1.76
2.	Iceland	(98)	(99)	35	1.81
3.	Azores I (São Miguel/Pico)	(81)	(100,101)	10	1.89
4.	Canaries (La Palma)	(102,103)	(102)	8	1.92
5.	French Polynesia	(104)	(104)	11	2.12
6.	Samoa	(105)	(105)	13	2.25
7.	Azores II (São Miguel)	(81)	(100)	13	2.33
8.	Réunion	(106)	(107)	(average)	2.42

\*Locality number used in main text Figure 3

†Number of individual data-points

‡ Average Pb model ages ( $t_m$ ) for each locality, see methods

Extended Data Table 2 | Input parameters for calculating Pb model ages

<i>Initial composition*</i> :	
<sup>206</sup> Pb/ <sup>204</sup> Pb ( <i>Canyon Diablo</i> ):	9.3066
<sup>207</sup> Pb/ <sup>204</sup> Pb ( <i>Canyon Diablo</i> ):	10.293
<sup>208</sup> Pb/ <sup>204</sup> Pb ( <i>Canyon Diablo</i> ):	29.475
μ <sub>1</sub> (1. Stage <sup>238</sup> U/ <sup>204</sup> Pb):	7.85
Th/U <sub>1</sub> (1. Stage):	3.876
t (time ago, Ga):	4.57
<sup>238</sup> U/ <sup>235</sup> U:	137.88†
k:	1.03326‡
<i>Decay constants (y<sup>-1</sup>):</i>	
λ <sub>238</sub> ( <sup>238</sup> U):	1.551E-10
λ <sub>235</sub> ( <sup>235</sup> U):	9.849E-10
λ <sub>232</sub> ( <sup>232</sup> Th):	4.948E-11

\*Radioactive element abundances and ratios are reported as present day values

†The “old consensus value” is used to be comparable with literature data

‡k is the conversion factor of Th/U weight ratios to atomic ratios of <sup>232</sup>Th/<sup>238</sup>U (or kappa). We use Th/U (weight ratios) throughout the text for consistency with most literature

# Promoterless gene targeting without nucleases ameliorates haemophilia B in mice

A. Barzel<sup>1</sup>, N. K. Paulk<sup>1</sup>, Y. Shi<sup>2</sup>, Y. Huang<sup>1</sup>, K. Chu<sup>1</sup>, F. Zhang<sup>1</sup>, P. N. Valdmanis<sup>1</sup>, L. P. Spector<sup>1</sup>, M. H. Porteus<sup>3</sup>, K. M. Gaensler<sup>2</sup> & M. A. Kay<sup>1</sup>

Site-specific gene addition can allow stable transgene expression for gene therapy. When possible, this is preferred over the use of promiscuously integrating vectors, which are sometimes associated with clonal expansion<sup>1</sup> and oncogenesis<sup>2</sup>. Site-specific endonucleases that can induce high rates of targeted genome editing are finding increasing applications in biological discovery and gene therapy<sup>3</sup>. However, two safety concerns persist: endonuclease-associated adverse effects, both on-target<sup>4</sup> and off-target<sup>5,6</sup>; and oncogene activation caused by promoter integration, even without nucleases<sup>7</sup>. Here we perform recombinant adeno-associated virus (rAAV)-mediated promoterless gene targeting without nucleases and demonstrate amelioration of the bleeding diathesis in haemophilia B mice. In particular, we target a promoterless human coagulation factor IX (*F9*) gene to the liver-expressed mouse albumin (*Alb*) locus. *F9* is targeted, along with a preceding 2A-peptide coding sequence, to be integrated just upstream to the *Alb* stop codon. While *F9* is fused to *Alb* at the DNA and RNA levels, two separate proteins are synthesized by way of ribosomal skipping. Thus, *F9* expression is linked to robust hepatic albumin expression without disrupting it. We injected an AAV8-*F9* vector into neonatal and adult mice and achieved on-target integration into ~0.5% of the albumin alleles in hepatocytes. We established that *F9* was produced only from on-target integration, and ribosomal skipping was highly efficient. Stable *F9* plasma levels at 7–20% of normal were obtained, and treated *F9*-deficient mice had normal coagulation times. In conclusion, transgene integration as a 2A-fusion to a highly expressed endogenous gene may obviate the requirement for nucleases and/or vector-borne promoters. This method may allow for safe and efficacious gene targeting in both infants and adults by greatly diminishing off-target effects while still providing therapeutic levels of expression from integration.

Site-specific gene targeting is one of the fastest growing fields in gene therapy and genome engineering. The rise in popularity of gene targeting can be attributed in large part to the development of readily engineered and easy to use site-specific endonucleases (for example, TAL- or CRISPR-based)<sup>3</sup> that can increase rates of gene disruption, gene correction or gene addition by as much as four orders of magnitude. However, these endonucleases may have significant adverse effects including immunogenicity, uncontrolled DNA damage response, off-target cleavage and mutagenesis, induction of chromosomal aberrations, as well as off-target integration of the transgene and endonuclease vectors (if DNA-based)<sup>4–6</sup>. When a vector-borne promoter, driving expression of the therapeutic transgene or and/or the nuclease, is integrated either on- or off-target, it may lead to undesired activation of nearby genes, including oncogenes. The use of endonucleases *in vivo* would require their vectorization, delivery and expression in a transient manner to minimize long-term side effects. It is unclear how integration of the vectored endonuclease gene could be strictly avoided.

Our promoterless, endonuclease-independent method harnesses the efficient transduction, favourable safety profile and high gene targeting rates associated with rAAV<sup>8–12</sup>, as well as the robust liver-specific

expression of the *Alb* locus<sup>13</sup>. Different rAAV serotypes can efficiently transduce various cell types *in vitro* or *in vivo*, while other serotypes have been designed or selected for desired phenotypes<sup>14–17</sup>. rAAV is in use in several clinical trials<sup>18,19</sup>. Notably, rAAV transduction allows high gene targeting rates *in vitro*<sup>8</sup> and *in vivo*<sup>9,20</sup>. The increased recombination rates may be due to the viral inverted terminal repeats, the encapsidation of single-stranded DNA, or the timing and subcellular localization of capsid uncoating.

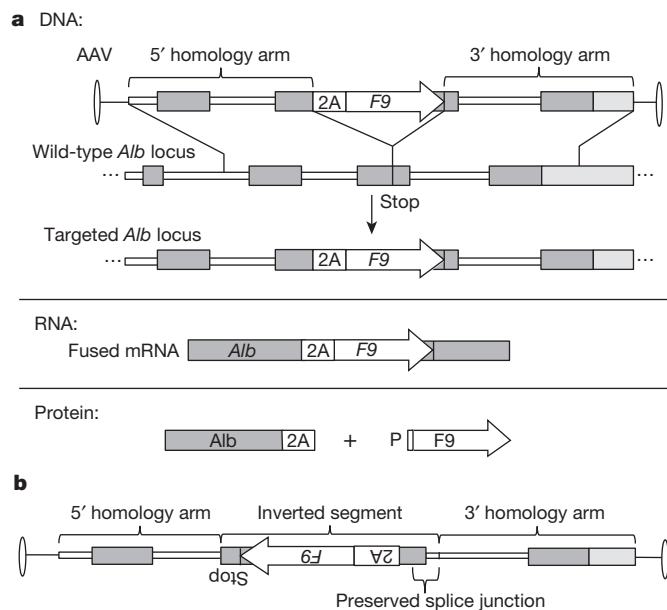
The safety of rAAV stems from its lack of pathogenicity, as well as being devoid of viral genes. Nevertheless, non-targeted genomic integration of rAAV occurs at a low but notable rate, and there are reports of such integrations inducing hepatocellular carcinoma in mice<sup>7</sup>. Transposition was attributed to vector integration at a chromosome 12 locus coding for imprinted genes and small RNAs. Integration of rAAV-borne promoters may be the leading cause of aberrant expression, as was established for lentiviral and retroviral vector integration leading to clonal expansion and oncogenesis<sup>1,21</sup>. Vector-borne promoters are in use in many continuing clinical trials. By contrast, the rAAV used in our strategy encodes no promoter, thus diminishing the chance of neighbouring oncogene activation in rare off-target integrations.

As proof of concept, we targeted the human *F9* gene, deficient in the X-linked recessive disease haemophilia B, which affects 1 in 30,000 males. Affected individuals suffer from serious spontaneous bleeding owing to a deficiency of plasma coagulation *F9* produced from the liver. Reconstitution with as little as 1–2% clotting factor can considerably improve quality of life, while 5–20% will markedly ameliorate the bleeding diathesis. Here we used the liver tropic rAAV8 serotype to target human *F9* for expression after integration from the robust liver-specific mouse *Alb* promoter. We postulated that: the *Alb* promoter should allow high levels of coagulation factor production even if integration takes place in only a small fraction of hepatocytes; and the high transcriptional activity at the *Alb* locus might make it more susceptible to transgene integration by homologous recombination.

Gene targeting without nucleases should affect only a small fraction of *Alb* alleles in the liver. Nevertheless, we opted to minimize disruption and dysregulation of the *Alb* gene by targeting *F9* as a 2A-fusion at the end of the *Alb* reading frame (Fig. 1a). 2A-peptides, derived from plus-strand RNA viruses, allow the production of several proteins from a single reading frame by means of ribosomal skipping<sup>22</sup>. This process leaves the first translated protein tagged with ~20 carboxy-terminal amino acids, and the second protein with just one additional amino-terminal proline. Functionality of both proteins is typically retained, and clinical trials using 2A-peptides did not report immunogenicity<sup>23</sup>. We used single-stranded AAV to target a codon-optimized *F9* coding sequence, preceded by a sequence coding for a porcine teschovirus-1 2A-peptide (P2A)<sup>22</sup>, to be integrated just 5' of the *Alb* stop codon. After integration, *Alb* and *F9* are co-transcribed from the strong *Alb* promoter, and should thus be co-regulated at the levels of splicing, nuclear exit, messenger RNA stability, translation initiation and endoplasmic reticulum localization. Two separate proteins are translated, both containing a signal

<sup>1</sup>Departments of Pediatrics and Genetics, 269 Campus Drive, CCSR Building, Room 2105, Stanford, California 94305-5164, USA. <sup>2</sup>Department of Medicine, Box 1270, UCSF, San Francisco, California 94143-1270, USA. <sup>3</sup>Department of Pediatrics, 269 Campus Drive, Lorry Lokey Stem Cell Research Building, Room G3045, Stanford, California 94305-5164, USA.



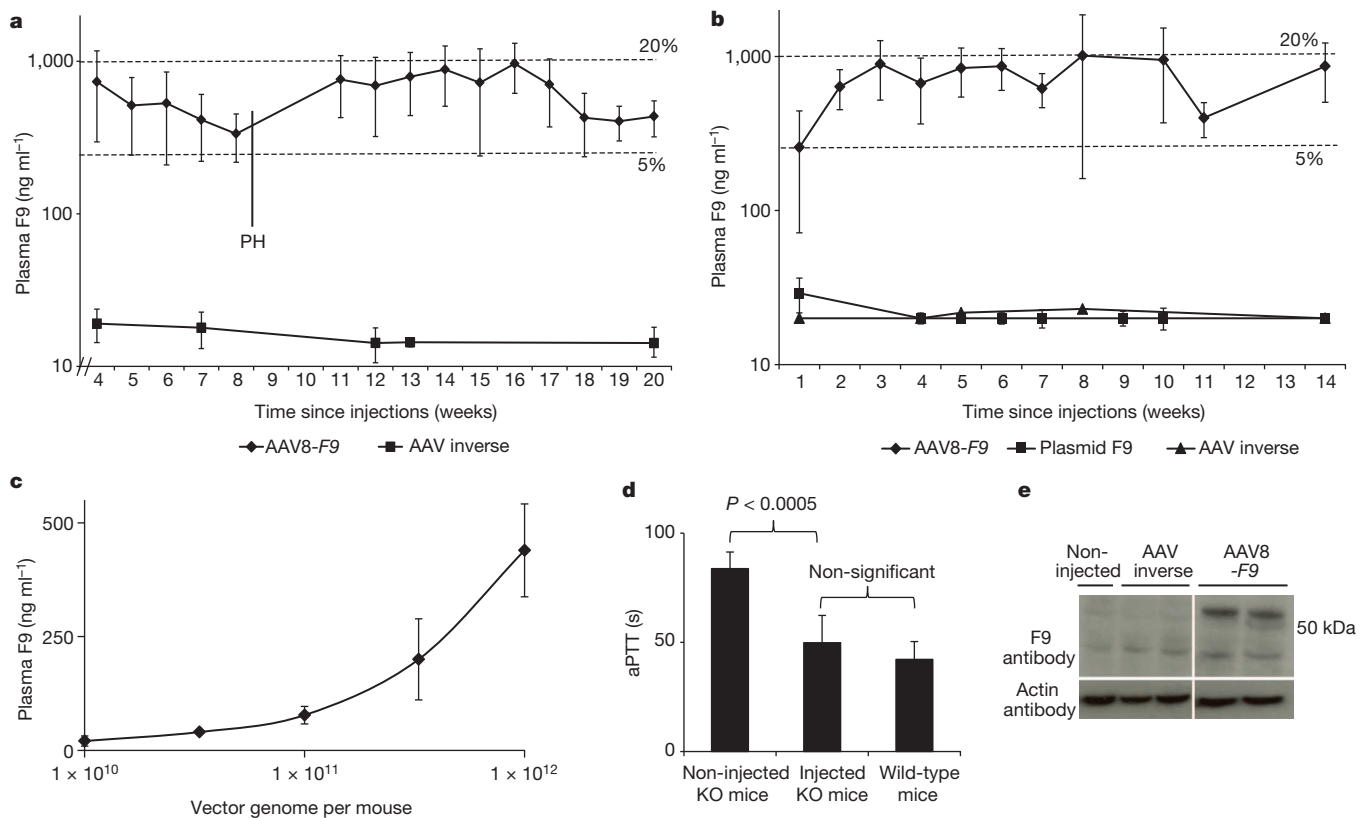


peptide, so that the endoplasmic reticulum-associated translation of *Alb* will be immediately followed by translation and processing of the clotting factor for secretion. Finally, to reduce the chance of off-target

**Figure 1 | Vector design and experimental scheme.** **a**, The rAAV8 vector encodes a codon-optimized human *F9* cDNA preceded by a 2A-peptide coding sequence and flanked by homology arms spanning the mouse *Alb* stop codon. Length of the 5' and 3' arms are 1.3 and 1.4 kb, respectively. After integration by homologous recombination, *Alb* and *F9* are fused at the DNA and RNA levels, but two separate proteins are produced as the result of ribosomal skipping. **b**, With respect to the *Alb* homology arms, the AAV inverse control has *F9* inverted along with the 2A-peptide coding sequence, the adjacent *Alb* exon and the preceding splice junction. Thin white lines denote *Alb* introns; dark grey boxes denote *Alb* exons; white boxes denote P2A; white arrows denote *F9* transgene; light grey boxes denote extragenic DNA. P, proline.

*F9* expression further, our vector has neither an ATG start codon before the *F9* signal peptide, nor a start codon in the 2A-peptide coding sequence or preceding *Alb* exon.

First, we performed intraperitoneal injections of 2-day-old C57BL/6 (B6) mice with  $2.5 \times 10^{11}$  vector genomes per mouse ( $\sim 1.25 \times 10^{14}$  per kg) of a rAAV8 coding for the human *F9* targeting cassette or a vector with an inverted cassette, controlling for off-target expression (Fig. 1b). The fragment inverted in the control with respect to the *Alb* homology arms includes not only the *F9* gene, but also the P2A coding sequence, the adjacent *Alb* exon, and the preceding splice junction. The inverse control should not allow notable *F9* expression after on-target integration, but would allow levels of off-target expression similar to that from the experimental construct. We measured plasma *F9* protein levels each week by enzyme-linked immunosorbent assay (ELISA), starting at 4 weeks of life (Fig. 2a). For the experimental group, levels of plasma



**Figure 2 | Human *F9* expression and activity in injected mice.** **a**, Plasma *F9* measured by ELISA following intraperitoneal injections of 2-day-old B6 mice with  $2.5 \times 10^{11}$  vector genomes per mouse of either the AAV8-*F9* experimental construct ( $n = 6$ ) or inverse control ( $n = 3$ ). The limit of detection was  $20 \text{ ng ml}^{-1}$ . PH, partial hepatectomy. Error bars represent s.d. Dashed lines denote 5% and 20% of normal *F9* levels. **b**, Plasma *F9* measured by ELISA after tail vein injections of 9-week-old female B6 mice with  $1 \times 10^{12}$  vector genomes per mouse of the AAV8-*F9* experimental construct ( $n = 7$ ), or inverse control ( $n = 3$ ), or a hydrodynamic injection of  $30 \mu\text{g}$  plasmid ( $3.5 \times 10^{12}$  copy number) coding for the *F9* construct in the 'correct'

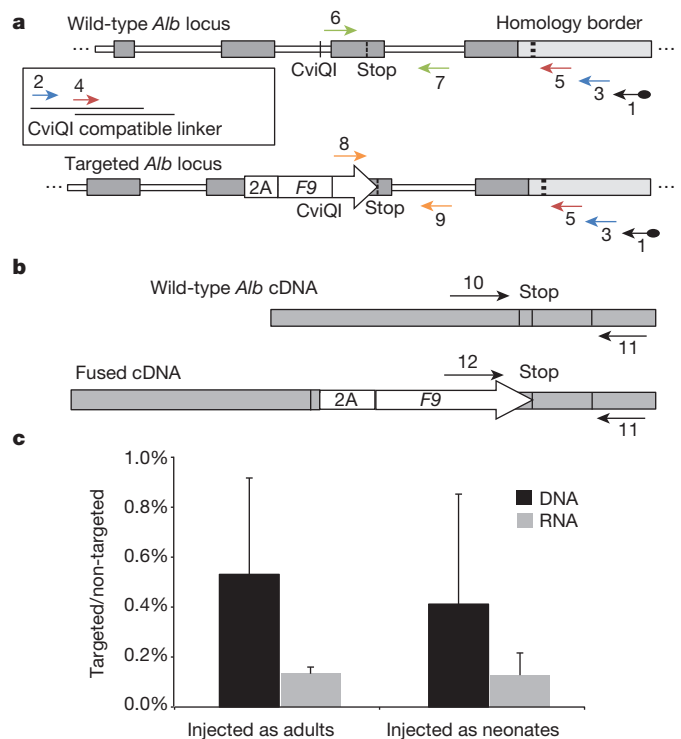
orientation ( $n = 3$ ). The limit of detection was  $20 \text{ ng ml}^{-1}$ . Error bars and dashed lines as in **a**. **c**, Plasma *F9* measured by ELISA following tail vein injections of 9-week-old female B6 mice with the designated vector dose of AAV8-*F9* experimental construct ( $n = 4$  for each dose group). Error bars represent s.d. **d**, Measurement of coagulation efficiency by activated partial thromboplastin time (aPTT) 2 weeks after tail vein injections of AAV8-*F9* at  $1 \times 10^{12}$  vector genomes per mouse ( $n = 5$ ). KO, knockout. Error bars represent s.d. **e**, Western blot analysis for *F9* in liver samples from mice injected with the AAV8-*F9* construct or inverse control. The expected size of human *F9* is 55 kilodaltons (kDa).

F9 plateaued at 350–1,000 ng ml<sup>-1</sup>, which corresponds to 7–20% of normal. For the inverse control group, F9 plasma levels were at or below the level of detection (20 ng ml<sup>-1</sup>), suggesting that in the experimental group, F9 expression does indeed originate from on-target integration. Importantly, F9 retains the original plasma protein levels after a two-thirds partial hepatectomy, a surgical procedure known to reduce episomal AAV transgene expression by >90%<sup>8</sup>, further establishing stable transgene integration.

To determine whether liver growth, as seen with neonates, is essential for therapeutic levels of gene targeting, we targeted *F9* to the *Alb* locus using the same vector in adult mice. Adult B6 mice were injected with  $1 \times 10^{12}$  vector genomes per mouse ( $\sim 5 \times 10^{13}$  vector genomes per kilogram) by tail vein with the AAV8-*F9* vector, or the inverse control. A third group of control mice received hydrodynamic tail vein injections of a plasmid coding for the promoterless *F9* construct in the 'correct' orientation. For the AAV8-*F9* mice group, plasma F9 levels were found to be stable at 7–20% of normal (Fig. 2b). Vector injections at lower dose led to lower plasma F9 levels without reaching a plateau at the doses tested (Fig. 2c). For adults as well as neonates, the F9 plasma levels of the inverse control group were at or below the limit of detection. Diminished F9 plasma levels were also associated with mice hydrodynamically injected with plasmid (Fig. 2b). Thus, targeting is dependent on rAAV vectorization. Finally, we performed rAAV injections in adult *F9* knockout haemophilia B mice. The functional coagulation, as determined by the activated partial thromboplastin time in treated knockout mice, was restored to levels similar to that of wild-type mice (Fig. 2d). The F9 biological activity correlated with plasma protein levels of  $709 \pm 91$  ng ml<sup>-1</sup>, similar to levels in wild-type mice (Fig. 2b–d).

F9 expression from the liver was confirmed by immunohistochemistry (Extended Data Fig. 1). Western blot analysis of liver samples detected F9 at the expected molecular mass, testifying that ribosomal skipping was efficient, and suggesting that both the ELISA and immunohistochemistry signals correspond to accurately processed F9 (Fig. 2d).

We opted to quantify targeting rates by quantitative PCR (qPCR). To avoid false signals from episomal rAAV, we first amplified a 3' segment of the genomic *Alb* locus in a manner not affected by presence or absence of an integrated *F9* sequence (Fig. 3a and Extended Data Fig. 2). The unbiased amplification was made possible by presence of a common restriction site at a roughly equal distance 3' of the stop codon in targeted and wild-type alleles. We then used the PCR amplicon as a template for two different qPCR assays: one quantifying the abundance of targeted *Alb* alleles, and the other quantifying the abundance of untargeted wild-type alleles. In the liver, only hepatocytes are targeted by rAAV8 (ref. 24). Therefore, we conservatively corrected for a 70% hepatocyte frequency<sup>25</sup> and found the rate of *Alb* alleles targeted by *F9* to be 0.5% on average for mice injected as either neonates or adults at the highest dose (Fig. 3c and Extended Data Fig. 3). Actual rates of targeting in neonates and adults might differ because AAV distribution to the liver may vary based on the different methods used for vector infusion. We then examined the proportion of fused *Alb-F9* mRNAs to wild-type *Alb* mRNAs by comparing two respective qPCR assays performed on an unbiased cDNA template (Fig. 3b). The proportion was found to be 0.1% on average for mice injected as either neonates or adults (Fig. 3c). This value tended to be lower than the rate of integration at the DNA level, although the difference was not statistically significant. It is possible that the production, processing and/or stability of chimaeric *Alb-F9* mRNA transcripts were reduced compared to wild-type *Alb* mRNA. While AAV8 has been shown to target only hepatocytes in the mouse liver<sup>24</sup>, here we did not rule out the possibility that some integration occurred in non-parenchymal cells that do not express albumin. Our observed targeting rate is higher than other reports<sup>9,12,20</sup>, and is particularly noteworthy in adult mice in which non-proliferating hepatocytes were expected to allow for a low rate of homologous recombination. We propose that the high expression rate at the *Alb* locus and the associated chromatin status may contribute to the high rates of targeting. Damage-induced proliferation cannot be strictly ruled out, but no



**Figure 3 | Rate of *Alb* targeting at the DNA and RNA levels.** **a**, Assessment of on-target integration rate begins using linear amplification with biotinylated primer 1 (black), annealing to the genomic locus but not to the vector. Linear amplicons are then bound to streptavidinylated beads and washed to exclude episomal vectors. Subsequent second-strand DNA synthesis with random primers was followed by CviQI restriction digestion. A compatible linker is then ligated, followed by two rounds of nested PCR (primers 2–3 in blue, and then primers 4–5 in red). CviQI cleaves at the same distance from the homology border in both targeted and wild-type alleles, thus allowing for unbiased amplification. The amplicons of the second nested PCR then serve as a template for qPCR assays with either primers 6–7 (green) or 8–9 (orange). **b**, For mRNA quantification, primers 10–11 or 11–12 were used to generate a cDNA for qPCR assays. Shape and fill code as in Fig. 1. **c**, Black bars represent the targeting rate of *Alb* alleles as the ratio between the abundance of the DNA template amplified by primers 6–7 to the abundance of the DNA template amplified by primers 8–9, corrected by a factor of 0.7 to account for hepatocyte frequency. Grey bars represent the expression rate of targeted *Alb* alleles as the ratio between the abundance of the cDNA template amplified by primers 10–11 to the abundance of the cDNA template amplified by primers 11–12.  $n = 3$  for each group, error bars represent s.d.

increase in alanine transaminase (ALT) levels was seen after injection (Extended Data Fig. 4).

AAV genomes may be present in cells as episomes, or as on- or off-target integrants. Total vector copy number was assessed by qPCR (Extended Data Fig. 5). The minor change in vector copy number after partial hepatectomy in mice injected as neonates may suggest that episomal vectors had already been greatly diluted during normal liver growth and development. In which case, vector copy number can be used as an approximate lower bound on the rate of off-target to on-target integration. However, most importantly, in the absence of a vector-borne promoter, *F9* should only be expressed from on-target integration. The reconstituted high F9 levels after partial hepatectomy (Fig. 2a) support this assumption, as only stably integrated transgenes could rebound after such a procedure, unlike that seen with transient episomal expression<sup>26</sup>. Lack of notable F9 plasma levels after treatment with the inverse control vector further demonstrated reduced off-target expression. We used quantitative reverse transcription PCR (qRT-PCR) to assess the ratio of fused *Alb-F9* mRNAs directly among the total *F9* mRNA pool (Fig. 4a). The ratio was found to be 1:1 for mice injected as neonates and as adults (Fig. 4b). This suggests that *F9* is expressed almost exclusively from

on-target integration. Indeed, the only specific signal from a northern blot with a P2A probe corresponded to the expected fused *Alb*-P2A-*F9* mRNA (Fig. 4c). Finally, a western blot with an anti-2A-peptide antibody indicated that the 2A-peptide is associated with a single species at the expected molecular mass of Alb (Fig. 4d), as would be expected only if expression was restricted to on-target integration and followed by efficient ribosomal skipping.

rAAV has become a popular vector for clinical therapy. Although the period of transgene expression in adults can last for several years, it is not yet known whether lifelong expression, as required for many genetic disorders, can be obtained with routine promoter-containing vectors. Episomal expression from AAV vectors is rapidly lost in dividing cells, even after just one round of cell division<sup>26</sup>. This makes it likely that diseases that induce regeneration and/or are treated in infancy while tissues continue to grow, will have limited durability. Secondary infusion of an AAV vector will be unlikely to result in a successful transduction, owing to the robust humoral immunity resulting from primary vector administration<sup>27</sup>. By contrast, our approach results in vector integration that would eliminate loss of expression over time, even in dividing tissues. This relies on the choice of an appropriate AAV serotype to avoid neutralization by pre-existing immunity.

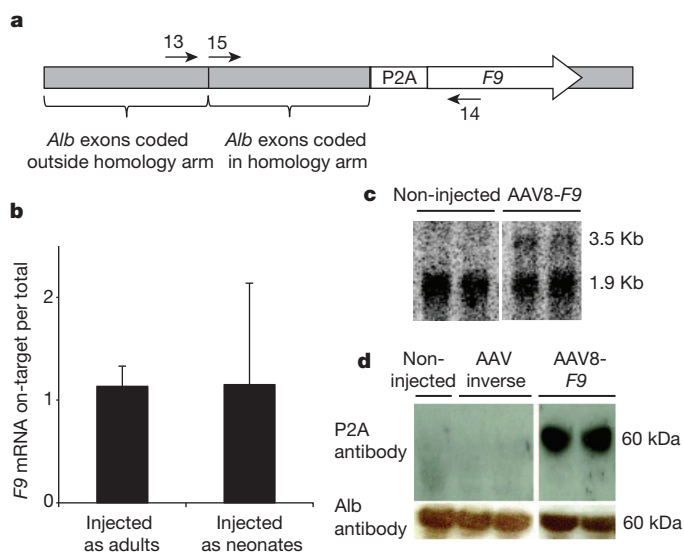
Previous work demonstrating targeting of *F9* to a chimaeric locus in a transgenic mouse<sup>10</sup> relied on the co-expression of nucleases that may be associated with immunological and genotoxic side effects. Interestingly, the same reliance on endonucleases held true even when *F8* was targeted to the *Alb* locus in mice and non-human primates<sup>28</sup>, probably because no homology arms were provided and integration relied instead on non-homologous end-joining. rAAV has already been used in clinical gene therapy trials to treat haemophilia B<sup>18</sup>. However, the transgene in these clinical trials was expressed from a vector-borne promoter that might induce oncogene activation, as has been reported in mice<sup>7</sup>. Measuring levels of alanine transaminases, we observed no liver toxicity with the injection of our human *F9* targeting vector (Extended Data Fig. 4). However, it remains to be determined whether the transgene overexpression associated with our method will lead to toxicity when

different therapeutic transgenes are targeted. 2A-induced immunogenicity could not be strictly ruled out, but notably no immune effects were reported in clinical trials when vector coding for a 2A peptide was targeted to lymphocytes<sup>23</sup>. Although we found no evidence of off-target expression and no ALT increase, the high vector dose we used may lead to other undesired outcomes such as increased off-target integration and increased immunogenicity. In the future, this could be mitigated by the use of AAV serotypes having better tropism and/or by use of hyperactive *F9* variants. Genetic polymorphisms at the target locus in the human patient population may lead to variable therapeutic efficacy owing to reduced homology. However, we found that ~95% of a 1000 Genomes Project (<http://www.1000genomes.org>) sample of the human population have no more than just two haplotypes at the relevant *ALB* sequence, which may enable broad applicability (Extended Data Table 1). Our work demonstrates a therapeutic effect for *in vivo* gene targeting without nucleases and without a vector-borne promoter. The favourable safety profile of our promoterless and nuclease-free gene targeting strategy for rAAV makes it a prime candidate for clinical assessment in the context of haemophilia and other genetic deficiencies<sup>29</sup>. More generally, this strategy could be applied whenever the therapeutic effect is conveyed by a secreted protein (for example, broadly neutralizing antibodies<sup>30</sup>) or when targeting confers a selective advantage<sup>12</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 April; accepted 12 September 2014.

Published online 29 October 2014.



**Figure 4 | Specificity of *F9* expression.** **a**, cDNA, produced from reverse transcription with a poly-dT primer, served as a template for a qPCR assay with either primers 13–14 or 14–15. **b**, Bars represent the rate of *Alb*-*F9* mRNAs to total *F9*-containing mRNAs as the ratio between the abundance of the cDNA template amplified by primers 13–14 to the abundance of the cDNA template amplified by primers 14–15.  $n = 3$  for each group, error bars represent s.d. **c**, Northern blot analysis of liver samples with a probe against P2A. The lower nonspecific signal corresponds in size to 18S rRNA. **d**, Western blot analysis of P2A from liver samples of mice injected with the AAV8-*F9* construct or inverse control. P2A is expected to be fused to albumin (66.5 kDa).

1. Aiuti, A. *et al.* Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. *Science* **341**, 1233151 (2013).
2. Hacein-Bey-Abina, S. *et al.* Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J. Clin. Invest.* **118**, 3132–3142 (2008).
3. Gaj, T., Gersbach, C. A. & Barbas, C. F. III. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31**, 397–405 (2013).
4. Hendel, A. *et al.* Quantifying genome-editing outcomes at endogenous loci with SMRT sequencing. *Cell Rep.* **7**, 293–305 (2014).
5. Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24**, 132–141 (2014).
6. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnol.* **31**, 822–826 (2013).
7. Donsante, A. *et al.* AAV vector integration sites in mouse hepatocellular carcinoma. *Science* **317**, 477 (2007).
8. Lisowski, L. *et al.* Ribosomal DNA integrating rAAV-rDNA vectors allow for stable transgene expression. *Mol. Ther.* **20**, 1912–1923 (2012).
9. Miller, D. G. *et al.* Gene targeting *in vivo* by adeno-associated virus vectors. *Nature Biotechnol.* **24**, 1022–1026 (2006).
10. Li, H. *et al.* *In vivo* genome editing restores haemostasis in a mouse model of haemophilia. *Nature* **475**, 217–221 (2011).
11. Shi, Y., Falahati, R., Zhang, J., Flebbe-Rehwal, L. & Gaensler, K. M. Role of antigen-specific regulatory CD4<sup>+</sup>CD25<sup>+</sup> T cells in tolerance induction after neonatal IP administration of AAV-hF.IX. *Gene Ther.* **20**, 987–996 (2013).
12. Paulk, N. K., Loza, L. M., Finegold, M. J. & Grompe, M. AAV-mediated gene targeting is significantly enhanced by transient inhibition of nonhomologous end joining or the proteasome *in vivo*. *Hum. Gene Ther.* **23**, 658–665 (2012).
13. Garcia-Martinez, R. *et al.* Albumin: pathophysiologic basis of its role in the treatment of cirrhosis and its complications. *Hepatology* **58**, 1836–1846 (2013).
14. Grimm, D. *et al.* *In vitro* and *in vivo* gene therapy vector evolution via multispecies interbreeding and retargeting of adeno-associated viruses. *J. Virol.* **82**, 5887–5911 (2008).
15. Lisowski, L. *et al.* Selection and evaluation of clinically relevant AAV variants in a xenograft liver model. *Nature* **506**, 382–386 (2014).
16. Li, C. *et al.* Single amino acid modification of adeno-associated virus capsid changes transduction and humoral immune profiles. *J. Virol.* **86**, 7752–7759 (2012).
17. Dalkara, D. *et al.* *In vivo*-directed evolution of a new adeno-associated virus for therapeutic outer retinal gene delivery from the vitreous. *Sci. Transl. Med.* **5**, 189ra176 (2013).
18. Nathwani, A. C. *et al.* Adenovirus-associated virus vector-mediated gene transfer in hemophilia B. *N. Engl. J. Med.* **365**, 2357–2365 (2011).
19. MacLaren, R. E. *et al.* Retinal gene therapy in patients with choroideremia: initial findings from a phase 1/2 clinical trial. *Lancet* **383**, 1129–1137 (2014).
20. Paulk, N. K. *et al.* Adeno-associated virus gene repair corrects a mouse model of hereditary tyrosinemia *in vivo*. *Hepatology* **51**, 1200–1208 (2010).
21. Cavazza, A., Moiani, A. & Mavilio, F. Mechanisms of retroviral integration and mutagenesis. *Hum. Gene Ther.* **24**, 119–131 (2013).
22. Kim, J. H. *et al.* High cleavage efficiency of a 2A peptide derived from porcine teschovirus-1 in human cell lines, zebrafish and mice. *PLoS ONE* **6**, e18556 (2011).



23. Johnson, L. A. *et al.* Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. *Blood* **114**, 535–546 (2009).
24. Nakai, H. *et al.* Unrestricted hepatocyte transduction with adeno-associated virus serotype 8 vectors in mice. *J. Virol.* **79**, 214–224 (2005).
25. Si-Tayeb, K., Lemaigre, F. P. & Duncan, S. A. Organogenesis and development of the liver. *Dev. Cell* **18**, 175–189 (2010).
26. Nakai, H. *et al.* Extrachromosomal recombinant adeno-associated virus vector genomes are primarily responsible for stable liver transduction *in vivo*. *J. Virol.* **75**, 6969–6976 (2001).
27. Calcedo, R. & Wilson, J. M. Humoral immune response to AAV. *Front. Immunol.* **4**, 341 (2013).
28. Anguela, X. M. *et al.* ZFN mediated targeting of albumin “safe harbor” results in therapeutic levels of human factor viii in a mouse model of hemophilia A. *Blood* **122**, 720 (2013).
29. Yew, N. S. & Cheng, S. H. Gene therapy for lysosomal storage disorders. *Pediatr. Endocrinol. Rev.* **11** (suppl. 1), 99–109 (2013).
30. Balazs, A. B. *et al.* Antibody-based protection against HIV infection by vectored immunoprophylaxis. *Nature* **481**, 81–84 (2011).

**Acknowledgements** This work was supported by a grant to M.A.K. from the National Heart Lung & Blood Institute (R01-HL064274). A.B. was supported by a fellowship from the Lucile Packard Foundation for Children’s Health, Stanford NIH-NCATS-CTSA UL1 TR001085 and Child Health Research Institute of Stanford University. N.K.P. was supported by a fellowship from the National Heart Lung & Blood Institute (F32-HL119059), the Hans Popper Memorial Fellowship from the American Liver Foundation, and the Stanford Dean’s Fellowship. M.H.P. was supported by the Laurie Krauss Lacob Faculty Scholar Fund in Pediatric Translational Medicine. The funding organizations played no role in experimental design, data analysis or manuscript preparation.

**Author Contributions** A.B., N.K.P., M.H.P., K.M.G. and M.A.K. designed the experiments. A.B., N.K.P., Y.S., Y.H., K.C., F.Z., P.N.V. and L.P.S. generated reagents and performed the experiments. A.B., N.K.P. and M.A.K. wrote and edited the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.A.K. ([markay@stanford.edu](mailto:markay@stanford.edu)).

## METHODS

**Plasmid construction.** A mouse genomic *Alb* segment (90474003–90476720 in NCBI reference sequence: NC\_000071.6) was PCR-amplified and inserted between AAV2 inverted terminal repeats into BsrGI and SpeI restriction sites in a modified pTRUF backbone<sup>31</sup>. The genomic segment spans 1.3 kilobases (kb) upstream and 1.4 kb downstream to the *Alb* stop codon. We then inserted into the Bpu10I restriction site an optimized P2A coding sequence preceded by a linker coding sequence (glycine–serine–glycine) and followed by an NheI restriction site. Finally, we inserted a codon-optimized F9 coding sequence into the NheI site to get pAB269 that served in the construction of the rAAV8 vector. To construct the inverse control, we first amplified an internal segment from the BsiWI restriction site to the 3' NheI restriction site. PCR primers used had 15 base tails to allow subsequent integration of the amplicon into a BsiWI and NheI cleaved plasmid at the inverse orientation using an In-Fusion Kit (Clontech). Primers used were: forward, 5'-ATGCAAGGCACGTACGTTATGTCAGCTTGGTCTTTCTTTGATCC-3', and reverse, 5'-TTTAGGCTAAGCTAGCTTTACTATGTCATTGCCTATGGCTATGAAGTG-3'. Final rAAV production plasmids were generated using an EndoFree Plasmid Megaprep Kit (Qiagen).

**rAAV vector production and titration.** rAAV8 vectors were produced as previously described using a  $\text{Ca}_3(\text{PO}_4)_2$  transfection protocol followed by CsCl gradient purification<sup>31</sup>. Vectors were titred by quantitative dot blot as described<sup>31</sup>.

**Mice injections and bleeding.** Animal work was performed in accordance to the guidelines for animal care at both Stanford University and the University of California San Francisco. Eight-week-old wild-type C57BL/6 (B6) mice were purchased from Jackson Laboratory to serve for adult injections and as breeding pairs to produce offspring for neonatal injections. Two-day-old wild-type B6 mice were injected intraperitoneally with  $2.5 \times 10^{11}$  vector genomes per mouse of rAAV8 (F9 or inverse) and bled weekly beginning at week 4 of life by retro-orbital bleeding for ELISA as previously described<sup>32</sup>. Adult (9-week-old) wild-type female B6 mice received either tail vein injections of rAAV8 (F9 or inverse, at the designated dose) or hydrodynamic injections of  $3.5 \times 10^{12}$  plasmid copies, and were similarly bled weekly for ELISA. F9 CD1/B6 hybrid knockout mice were injected as adults with  $1 \times 10^{12}$  vector genomes per mouse of rAAV8 F9, and bled retro-orbitally two weeks after injection for ELISA and activated partial thromboplastin time assays, as previously described<sup>11</sup>.

**ELISA.** ELISA for F9 was performed as previously described<sup>32</sup> with the following antibodies: mouse anti-human F9 IgG primary antibody at 1:1,000 (Sigma F2645), and polyclonal goat anti-human F9 peroxidase-conjugated IgG secondary antibody at 1:4,200 (Enzyme Research GAFIX-APHRP).

**Partial hepatectomies.** The two-thirds partial hepatectomies were performed as previously described<sup>33</sup> with the following minor modifications: no surgical retractors were used to maintain an open abdominal cavity, 3-0 Sof silk wax coated braided silk suture (Covidien S-184) was used for knotting desired lobes for resection, and 6-0 Polysorb polyester suture (Covidien GL889CV11) was used for closing the peritoneum and abdominal skin.

**Activated partial thromboplastin time assay.** The activated partial thromboplastin time assay was carried out using a SCA2000 veterinary coagulation analyser (Synbioics) according to manufacturer's instructions<sup>11</sup>.

**Western blot analysis.** Western blots for the detection of F9, albumin, P2A and actin used the following antibodies: polyclonal goat anti-human F9 peroxidase-conjugated IgG primary antibody at 1:20,000 (Enzyme Research GAFIX-APHRP), polyclonal rabbit anti-mouse serum albumin IgG primary antibody at 1:40,000 (Abcam ab19196), donkey anti-rabbit peroxidase-conjugated IgG secondary antibody at 1:10,000 (ECL NA-9340), polyclonal rabbit anti-2A peptide primary antibody at 1:10,000 (Millipore ABS31), and monoclonal mouse anti- $\beta$ -actin peroxidase-conjugated IgG primary antibody at 1:50,000 (Sigma A3854).

**Northern blot analysis.** Northern blots for the detection of P2A-F9 coding mRNAs used the following <sup>32</sup>P end-labelled anti-2A probe: 5'-GCCAGGTTCTCTTCCACGTCGCCGCTGTTTCAGCAGGCTGAAATTGGTGGCGCCGCT-3'.

**Assessing rate of *Alb* locus targeting by qPCR.** Amplification of desired genomic *Alb*, but not undesired vector amplification, began by performing linear amplification with the following dual-biotinylated primer: 5'-/2-biotin/GTCTCTCATTCAGAAATCTCGTAATGTTGAAG-3', annealing outside the arm of homology. Subsequent second-strand DNA synthesis was followed by CviQI restriction digestion to produce fragments of roughly equal size and known sticky ends from both targeted and wild-type alleles. Two oligonucleotides were annealed to a CviQI-compatible linker. Oligonucleotide sequences were 'Watson': 5'-CTGAAGGCTCAGGTTACACAGGCAGCTCGTAGGAGGTGTTCCAGTTCACACG-3', and 'Crick': 5'-TACGTGGTGAAGTGAACACCTCCTACGAGC/3ddC/-3'. Linker ligation was followed by a primary nested PCR, using primers 2: 5'-CTGAAGGCTCAGGTACACAGGCAC-3', and 3: 5'-GTATTGGTTTCTAGGGTACACCCATAAG-3', and a second nested PCR using primers 4: 5'-GCTCGTAGGAGGTGTTCCAGTTCAC-3', and 5: 5'-GGGAGAGTATTAACGTTTATTTTCATTGTGTT

G-3'. No nested PCR product was detected in a control without linear amplification. The PCR amplicon served as a template for two different TaqMan qPCR assays. To quantify the abundance of wild-type *Alb* alleles, we used a TaqMan qPCR with primers 6: 5'-TGCCTATGGCTATGAAGTGC-3', and 7: 5'-CTGAGAAGGTTGTGGTTGTGA-3', and TaqMan probe: 5'-TGCAAAGACGCCTTAGCCTAACACA-3'. To quantify the abundance of targeted alleles we used a qPCR with primers 8: 5'-GATACGTGAAGTGGATCAAAGAAA-3' and 9: 5'-CAAATGGTTATCAGTCTTGATCG-3' and TaqMan probe: 5'-CACATCACACCAACCAACCTTCTCAGGT-3'. For non-injected controls, no qPCR signal was detected with primers 8–9. The abundance of the template for each amplicon was calculated using its own standard curve. We calculated the ratio between the abundance of the template for the 6–7 primer pair to the abundance of the template for the 8–9 primer pair. We then conservatively corrected this ratio by multiplying it by a factor of 0.7 to account for hepatocyte frequency in the samples<sup>25</sup>, as only hepatocytes are targeted by rAAV8 in the liver<sup>24</sup>.

**Assessing rate of F9-containing *Alb* mRNAs by qPCR.** cDNA produced from reverse transcription with a poly-dT primer served as a template for two different TaqMan qPCR assays. We quantified the abundance of wild-type *Alb* mRNA by qPCR with primers 10: 5'-CTGACAAGGACACCTGCTTC-3' and 11: 5'-TGAGTCCTGAGTCTTCATGTCTT-3', and TaqMan probe: 5'-CCACAACCTTCTCAGGCTACCTGA-3'. We quantified the abundance of fused mRNAs by a TaqMan qPCR with primers 12: 5'-CCAAGGTGTCAGATACGTG-3' and 11: 5'-TGAGTCTGCTGAGTCTTCATGTCTT-3', and TaqMan probe: 5'-CCACAACCTTCTCAGGCTACCTGA-3'. For non-injected controls, no qPCR signal was detected with primers 11–12. For the no-reverse-transcriptase control, no signal was detected with primers 10–11 and 11–12. The rate of F9-containing *Alb* mRNAs was calculated as the ratio between the abundance of template for the 10–11 primer pair to the abundance of template for the 11–12 primer pair.

**Assessing specificity of F9 expression.** cDNA produced from reverse transcription with a poly-dT primer served as a template for two different TaqMan qPCR assays. We quantified the abundance of on-target F9 expression using a qPCR with primers 13: 5'-CTTGGGCTTGCTTCAC-3' and 14: 5'-AGGATCTTGTGGCGTTCTC-3', and TaqMan probe: 5'-CGGTACACTCGGCGCTCAGC-3'. We quantified total F9-containing mRNA abundance using a qPCR with primers 15: 5'-GCTCTTGCTGAGCTGGTGA-3' and 14: 5'-AGGATCTTGTGGCGTTCTC-3', and TaqMan probe: 5'-CGACTGAGGCTCCAAACCTTGTC-3'. Calculations of abundance used a separate standard curve for each pair. No qPCR signal was detected for no-reverse transcriptase and non-injected controls. The rate of *Alb*-F9 mRNAs to total F9-containing mRNAs was determined by comparing the abundance of template for the 13–14 primer pair to abundance of template for the 14–15 primer pair.

**Assessing vector copy number.** DNA purified from mice liver served as a template for two different TaqMan qPCR assays. We quantified the abundance of haploid mouse genomes using a qPCR with primers 'Alb copy number Ref F': 5'-CAGAACGGTCTTTCTCGGAT-3' and 'Alb copy number Ref R': 5'-TCTTCATCCTGCCCTAAACC-3', and TaqMan probe: 5'-CTCAGCCCTGCAGTGTGCA-3'. We quantified the abundance of AAV genomes (episomal or integrated) using qPCR with primers 8: 5'-GATACGTGAAGTGGATCAAAGAAA-3' and 9: 5'-CAAAATGGTTATCAGTCTTGATCG-3', and TaqMan probe: 5'-CACATCACAAACCAACCTTCTCAGGT-3'. For non-injected controls, no qPCR signal was detected with primers 8–9. The abundance of the template for each amplicon was calculated using its own standard curve. Vector copy number per haploid genome was calculated as the ratio between the abundance of the template for the 8–9 primer pair to the abundance of the template for the Alb copy number Ref F and Alb copy number Ref R primer pair.

**Statistics.** Statistical analyses were conducted with Microsoft Excel. Experimental differences were evaluated by a Student's one-tailed *t*-test assuming equal variance (except for the ALT measurement, where equal variance was not assumed).

**Liver immunohistochemistry.** Liver lobes were collected from mice, rinsed in ice cold dPBS, blotted dry, mounted in OCT (Tissue-Tek 4583) and frozen in cryomolds (Tissue-Tek 4557) in a liquid nitrogen cooled methylbutane bath. Cryomolds were placed at  $-80^\circ\text{C}$  until sectioning. In all samples, a minimum of two liver sections from varying depths were cut at 5  $\mu\text{m}$  and mounted onto positively charged glass slides (Thermo 6776214). Fluorescent staining was performed according to established protocols<sup>34</sup> with minor modifications. In brief, frozen sections were thawed to room temperature, fixed in acetone for 10 min, removed and allowed to air dry. Slides were then washed three times in dPBS for 2 min, blocked with 5% donkey serum (Santa Cruz sc-2044) for 15 min at room temperature in a humidified chamber, washed once in dPBS for 1 min and then sections were circled with an ImmEdge pen (Vector Laboratories H-4000). Slides were incubated with 200  $\mu\text{l}$  of goat anti-human F9 IgG primary antibody (Affinity Biologicals, GAFIX-AP) at 1:100 in 5% donkey serum for 3 h and then washed three times in dPBS for 2 min. Slides were then incubated with 200  $\mu\text{l}$  of donkey anti-goat AlexaFluor 594 IgG secondary

antibody (Life Technologies A11058) at 1:400 in dPBS for 30 min and then washed three times in dPBS for 2 min. Slides were rinsed with distilled water and mounted with 80 µl ProLong Gold Antifade with DAPI nuclear counterstain (Life Technologies P36935) and covered with a #1.5 coverslip (Thermo 12-544-G). Fluorescent images were taken on a Zeiss Observer.Z1 microscope equipped with a Zeiss AxioCam MRc colour camera and Zeiss AxioVision software (version 4.8.2.0). Images were overlaid using Adobe Photoshop CS6 software (version 13 x64). Controls included no-primary secondary-only antibody staining, and comparison to positive control frozen human liver tissue sections (Zyagen HF-314) and negative control frozen untreated mouse liver sections.

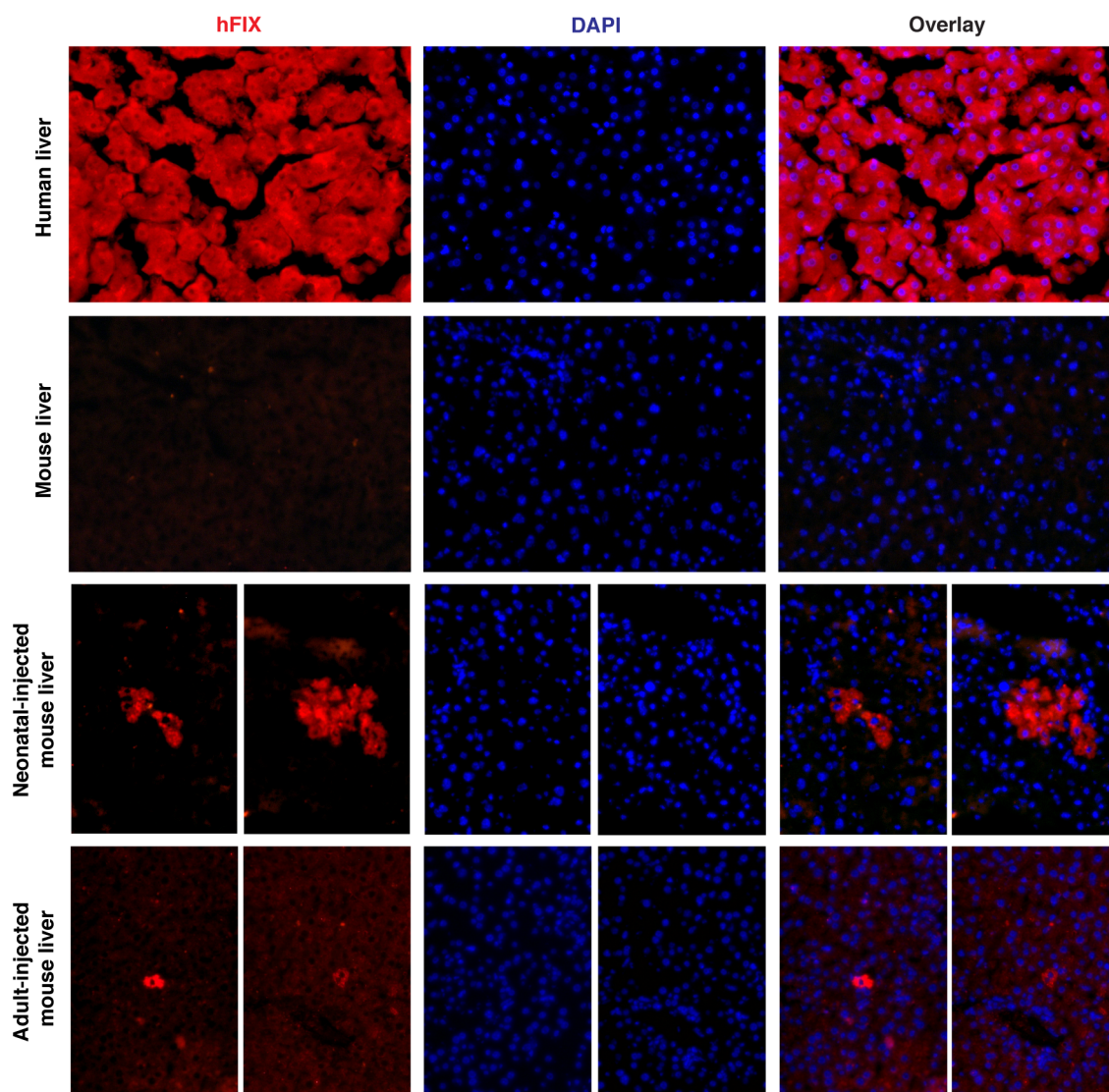
**ALT measurements.** Serum ALT measurements were performed on mouse serum obtained via retro-orbital bleeding using an ALT kinetic measurement kit (Teco Diagnostics) compared with a standard curve. AAV8-U6 and H1 promoter short hairpin RNA (shRNA) sequences are derived from on shRNA toxicity studies performed previously<sup>35</sup>.

**Assessing the distribution of *Alb* haplotypes in the human population.** A selection of the ShapeIt2 phased haplotypes for the 1000 Genomes Phase 1 integrated variant calls, corresponding to a region 1.3-kb upstream and 1.4-kb downstream from the human *F9* integration site at the *Alb* stop codon, was downloaded from

[ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis\\_results/shapeit2\\_phased\\_haplotypes/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/shapeit2_phased_haplotypes/) using the 1000 Genomes Data Slicer tool (available at [http://browser.1000genomes.org/Homo\\_sapiens/UserData/SelectSlice](http://browser.1000genomes.org/Homo_sapiens/UserData/SelectSlice)). Haplotypes consisting of single nucleotide polymorphisms with a substantial frequency of the alternative allele in all populations were combined (here, greater than or equal to 45%, whereas those excluded were less than 1%) and treated as individual strings for calculating population frequency.

31. Grimm, D., Pandey, K., Nakai, H., Storm, T. A. & Kay, M. A. Liver transduction with recombinant adeno-associated virus is primarily restricted by capsid serotype not vector genotype. *J. Virol.* **80**, 426–439 (2006).
32. Lu, J., Zhang, F. & Kay, M. A. A mini-intronic plasmid (MIP): a novel robust transgene expression vector *in vivo* and *in vitro*. *Mol. Ther.* **21**, 954–963 (2013).
33. Mitchell, C. & Willenbring, H. A reproducible and well-tolerated method for 2/3 partial hepatectomy in mice. *Nature Protocols* **3**, 1167–1170 (2008).
34. Rogers, G. L. & Hoffman, B. E. Optimal immunofluorescent staining for human factor ix and infiltrating T cells following gene therapy for hemophilia B. *J. Genet. Syndr. Gene Ther. Suppl.* **1**, 012 (2012).
35. Grimm, D. *et al.* Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. *Nature* **441**, 537–541 (2006).

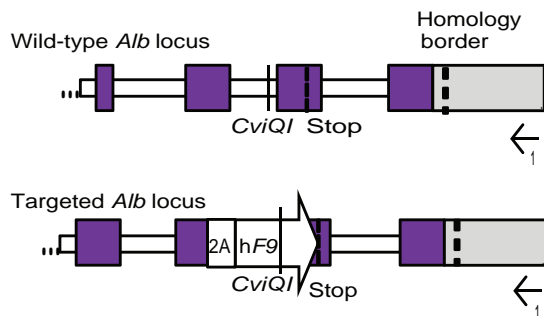




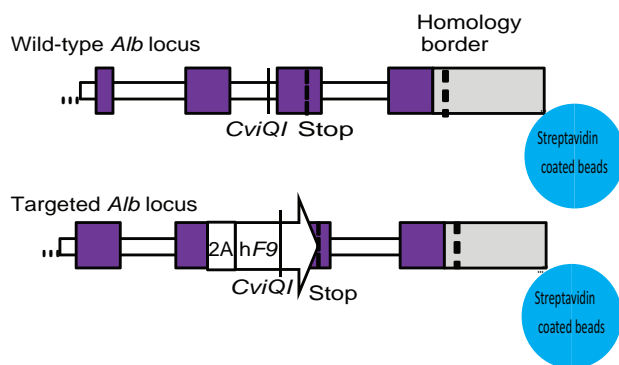
**Extended Data Figure 1 | Human F9 liver immunohistochemistry.** From top to bottom, panels show human F9 staining (red) with 4',6-diamidino-2-phenylindole (DAPI) nuclear counterstain (blue) in positive control human

liver, negative control untreated mouse liver, and two sets of representative stains from mice treated as neonates or adults with AAV8-F9. Original magnification,  $\times 200$ .

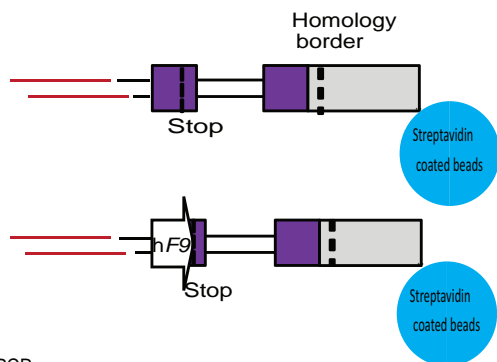
## Step 1: Linear amplification (LAM)



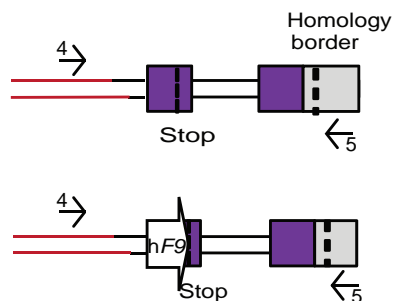
## Step 3: Second strand DNA synthesis with random primers



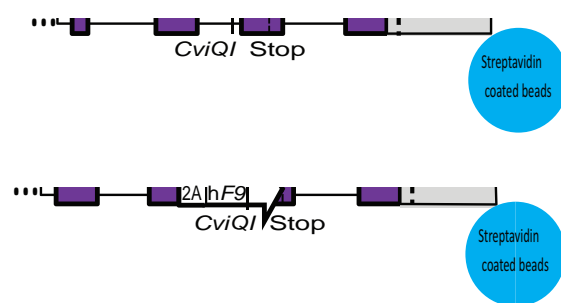
## Step 5: Linker ligation



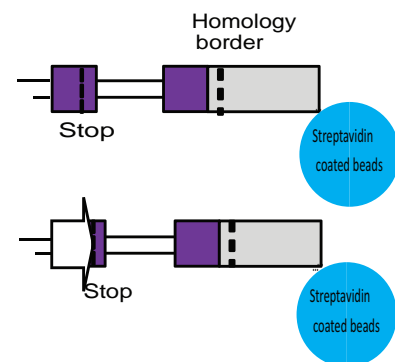
## Step 7: 2nd nested PCR



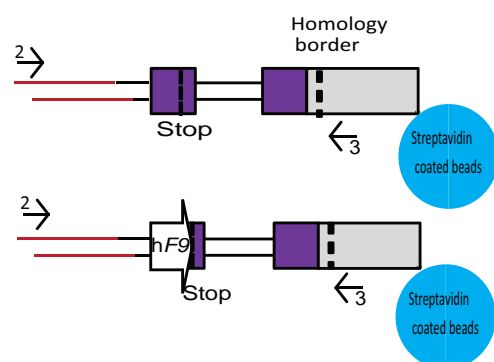
## Step 2: Binding to beads and wash



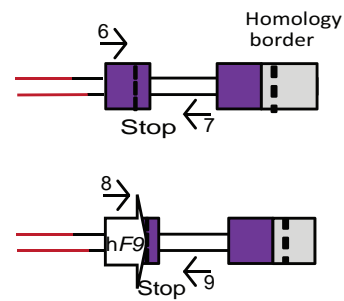
## Step 4: CviQI cleavage



## Step 6: 1st nested PCR

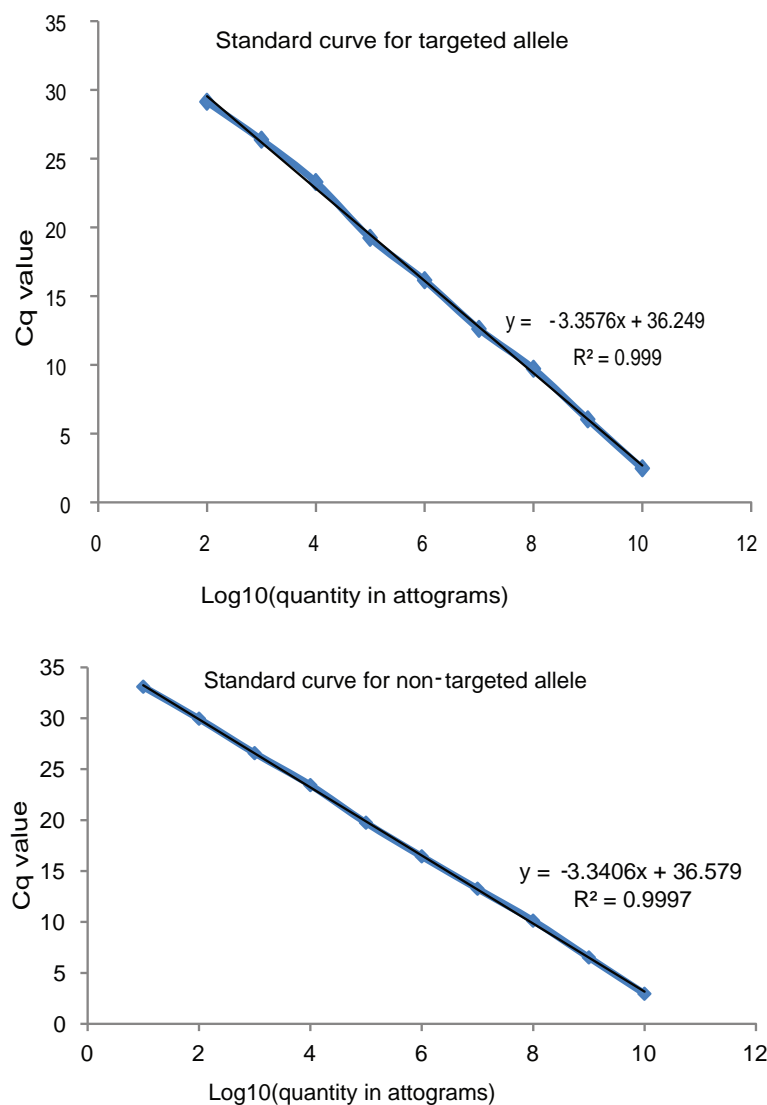


## Step 8: qPCR



**Extended Data Figure 2 | Scheme of targeting rate assessment.** Assessment of on-target integration rate begins using linear amplification with biotinylated primer 1 (black), annealing to the genomic locus but not to the vector (step 1). Linear amplicons are then bound to streptavidinylated beads and washed to exclude episomal vectors (step 2). Subsequent second-strand DNA synthesis with random primers (step 3) was followed by CviQI restriction digestion

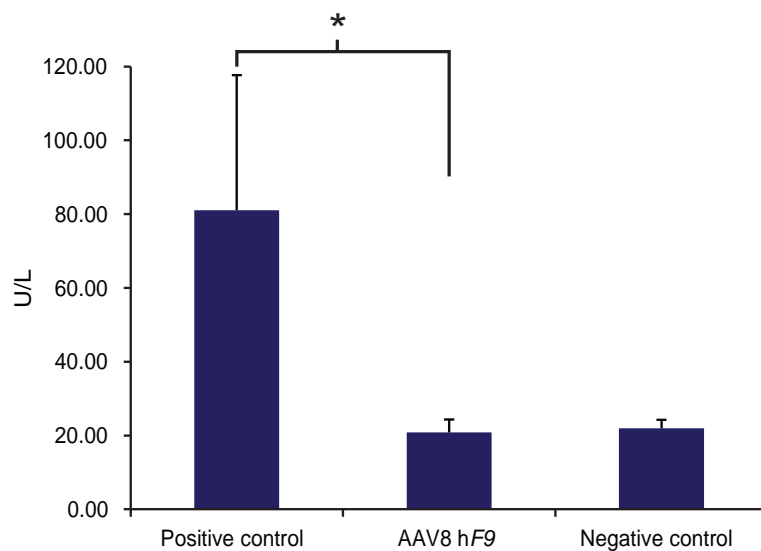
(step 4). A compatible linker is then ligated (step 5) followed by two rounds of nested PCR amplifications (primers 2–3 in blue (step 6), and then primers 4–5 in red (step 7)). CviQI cleaves at the same distance from the homology border in both targeted and wild-type alleles, thus allowing for unbiased amplification. The amplicons of the second nested PCR then serve as a template for qPCR assays with either primers 6–7 (green) or 8–9 (orange) (step 8).



**Extended Data Figure 3 | Standard curves for targeting rate assessment by qPCR.** qPCR standard curves for the targeted allele (primers 8 and 9, Fig. 3) and non-targeted allele (primers 6 and 7, Fig. 3). Mass units used

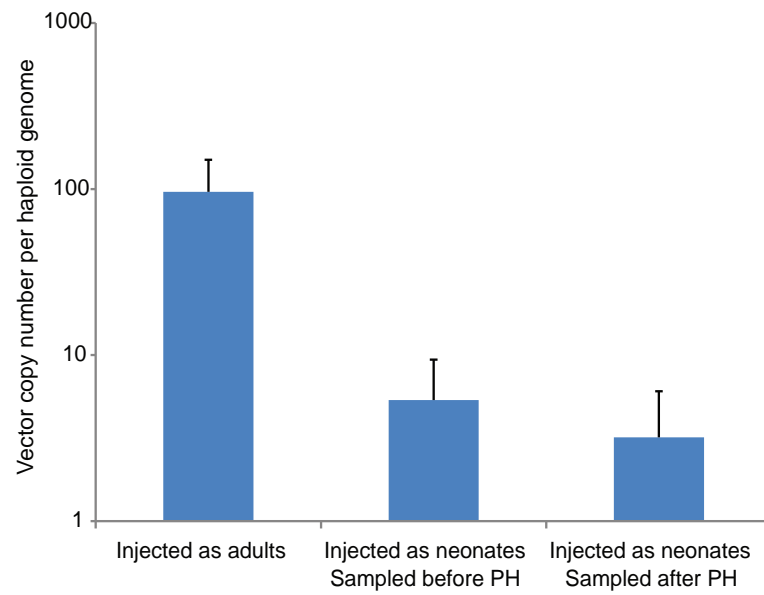
are functionally equivalent to molarity because all amplicons used were of equal length.



**Extended Data Figure 4 | Toxicity assessment by ALT measurement.**

Alanine transaminase levels (ALT) were evaluated 7 days after injection in mice injected with AAV8 coding for our experimental vector ( $1 \times 10^{12}$ ) or a negative control coding for a known non-toxic cassette ( $1 \times 10^{12}$  of H1 promoter-driven shRNA), or a positive control coding for a known toxic

cassette ( $5 \times 10^{11}$  of U6 promoter-driven shRNA). Data represent mean of two measurements of four independent mice for each groups. The statistical significance is defined here as having  $P < 0.05$  in a one-tailed  $t$ -test between samples of different variance.



**Extended Data Figure 5 | Vector copy number.** Vector copy number assessed by qPCR using primers 8 and 9 (Fig. 3).  $n = 7$  for mice injected as adults;  $n = 6$  for mice injected as neonates and analysed before or after partial hepatectomy (PH). Error bars represent s.d.

Extended Data Table 1 | Haplotypes in the human population at the relevant *ALB* locus as extracted from the 1000 Genomes Project

Position	ID	REF	ALT	HAP1	HAP2	HAP3	HAP4	HAP5	HAP6	HAP7
74285239	rs962004	C	T	T	C	C	C	C	T	C
74285552	rs4076	A	G	A	G	A	A	A	G	G
74285567	rs962005	C	A	C	A	C	C	C	A	A
74285758	rs2236766	G	T	G	T	T	G	G	T	T
74285823	rs2236767	G	A	G	A	G	G	G	A	A
74287403	rs4429703	T	C	C	T	T	C	T	T	C
Frequency				50.14%	44.69%	2.15%	1.51%	1.37%	0.09%	0.05%



# Productivity limits and potentials of the principles of conservation agriculture

Cameron M. Pittelkow<sup>1\*†</sup>, Xinqiang Liang<sup>2\*</sup>, Bruce A. Linquist<sup>1</sup>, Kees Jan van Groenigen<sup>3</sup>, Juhwan Lee<sup>4</sup>, Mark E. Lundy<sup>1</sup>, Natasja van Gestel<sup>3</sup>, Johan Six<sup>4</sup>, Rodney T. Venterea<sup>5,6</sup> & Chris van Kessel<sup>1</sup>

One of the primary challenges of our time is to feed a growing and more demanding world population with reduced external inputs and minimal environmental impacts, all under more variable and extreme climate conditions in the future<sup>1–4</sup>. Conservation agriculture represents a set of three crop management principles that has received strong international support to help address this challenge<sup>5,6</sup>, with recent conservation agriculture efforts focusing on smallholder farming systems in sub-Saharan Africa and South Asia<sup>7</sup>. However, conservation agriculture is highly debated, with respect to both its effects on crop yields<sup>8–10</sup> and its applicability in different farming contexts<sup>7,11–13</sup>. Here we conduct a global meta-analysis using 5,463 paired yield observations from 610 studies to compare no-till, the original and central concept of conservation agriculture, with conventional tillage practices across 48 crops and 63 countries. Overall, our results show that no-till reduces yields, yet this response is variable and under certain conditions no-till can produce equivalent or greater yields than conventional tillage. Importantly, when no-till is combined with the other two conservation agriculture principles of residue retention and crop rotation, its negative impacts are minimized. Moreover, no-till in combination with the other two principles significantly increases rainfed crop productivity in dry climates, suggesting that it may become an important climate-change adaptation strategy for ever-drier regions of the world. However, any expansion of conservation agriculture should be done with caution in these areas, as implementation of the other two principles is often challenging in resource-poor and vulnerable smallholder farming systems, thereby increasing the likelihood of yield losses rather than gains. Although farming systems are multifunctional, and environmental and socio-economic factors need to be considered<sup>14–16</sup>, our analysis indicates that the potential contribution of no-till to the sustainable intensification of agriculture is more limited than often assumed.

To help address global food security challenges, conservation agriculture holds much promise as ‘an approach to managing agro-ecosystems for improved and sustained productivity, increased profits and food security while preserving and enhancing the resource base and the environment’<sup>13</sup>. Conservation agriculture represents a set of three crop management principles: (1) direct planting of crops with minimum soil disturbance (that is, no-till), (2) permanent soil cover by crop residues or cover crops, and (3) crop rotation<sup>5,6</sup>. In recent decades, widespread adoption of no-till has occurred over approximately 125 million hectares, equivalent to 9% of global arable land, with varying degrees of application of the other two conservation agriculture principles<sup>5,13</sup>. However, the impacts of no-till by itself and conservation agriculture on crop productivity remain contested<sup>5–13</sup>. Here, we synthesized current scientific evidence at a global scale to assess crop yields under no-till in relation to implementation of the other two conservation agriculture principles, residue retention and crop rotation.

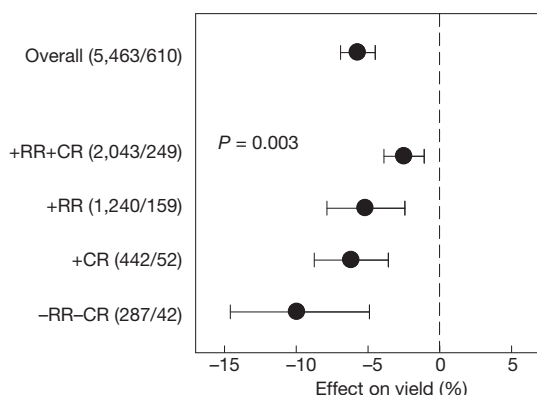
A comprehensive meta-analysis was performed on data from peer-reviewed publications, representing the largest assessment so far on this topic. Because not all three principles of conservation agriculture are adopted by all farmers<sup>8,17</sup>, studies at a minimum had to include no-till, the original and central concept of conservation agriculture, and conventional tillage treatments (note: minimum-tillage practices were not included). Only field experiments containing side-by-side yield comparisons were included in the database (see Methods for study selection details). Since conservation agriculture is not necessarily a low-input form of agriculture and in fact has been adopted to the greatest extent<sup>13</sup> in countries characterized by highly mechanized, high-input agricultural systems, all comparisons were included regardless of input intensity. To examine how the effects of no-till changed across the other two principles of conservation agriculture, yield comparisons were grouped into categories based on the presence or absence of residue management and crop rotation practices as determined by information reported in the original studies. For each paired yield comparison, no-till and conventional tillage treatments received the same residue management and rotation practices. In total, the database consisted of 5,463 observations from 610 studies.

Overall, we found that no-till negatively impacts crop yields by 5.7% (Fig. 1), although under certain conditions it produces yields equivalent to or even greater than conventional tillage systems (Figs 2 and 3). To limit global agricultural expansion and thereby reduce net environmental degradation, enhancing production per unit area through agricultural intensification efforts has been identified as a promising approach<sup>3,4,14,18</sup>. However, our meta-analysis indicates that no-till is limiting rather than enhancing global crop production and sustainable intensification efforts. Certainly, yield is only one component of agricultural systems, and there is an urgent need to optimize farming practices across other environmental and socio-economic performance indicators<sup>1,15</sup>. We recognize that in many, but definitely not all situations, continuous no-till along with the other two conservation agriculture principles may represent a more profitable management system (often because of reduced energy/diesel costs related to tillage), with the potential to improve soil quality and provide greater ecosystem services<sup>16,17,19</sup>. In addition, as agricultural crop yields are variable in time and space, yield outcomes can be difficult to predict at the individual farm-scale.

Importantly, the negative impacts of no-till are minimized when both of the other conservation agriculture principles are also applied (–2.5%) (Fig. 1). The largest yield declines occur when no-till is implemented alone (–9.9%) or with only one other conservation agriculture principle (–5.2 and –6.2% for residue retention and crop rotation, respectively). To help close the yield gap with conventional tillage, these findings suggest that instead of implementing no-till as the first step towards conservation agriculture in cropping systems where residue retention and crop rotation are absent (and anticipating that these two principles will follow in

<sup>1</sup>Department of Plant Sciences, University of California, Davis, California 95616, USA. <sup>2</sup>College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China. <sup>3</sup>Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, Arizona 86011, USA. <sup>4</sup>Department of Environmental Systems Science, Swiss Federal Institute of Technology, ETH-Zurich, Zurich 8092, Switzerland. <sup>5</sup>United States Department of Agriculture, Agricultural Research Service, Soil and Water Management Unit, St Paul, Minnesota 55108, USA. <sup>6</sup>Department of Soil, Water, and Climate, University of Minnesota, St Paul, Minnesota 55108, USA. <sup>†</sup>Present address: Department of Crop Sciences, University of Illinois, Urbana, Illinois 61801, USA.

\*These authors contributed equally to this work.



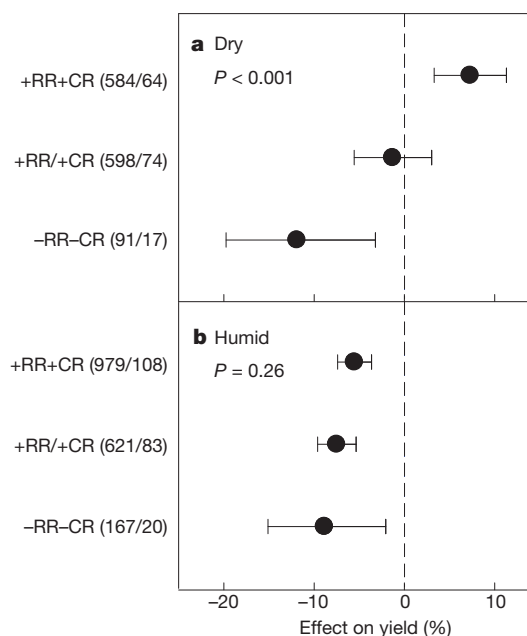
**Figure 1 | Comparison of yield in no-till versus conventional tillage systems in relation to the other two principles of conservation agriculture.**

Results are shown for the entire data set (overall) and for subcategories of studies which indicated the presence or absence of residue retention and crop rotation for both no-till and conventional tillage systems: +RR+CR (residue retention + crop rotation), +RR (residue retention), +CR (crop rotation), or -RR-CR (without residue retention or crop rotation). The number of observations and total number of studies included in each category are displayed in parentheses. Error bars represent 95% confidence intervals. Significant differences between categories are indicated by *P* values based on randomization tests.

time), the primary focus should be on implementing no-till systems that already employ the other two principles. This conclusion has important implications for the promotion of conservation agriculture as an agricultural development strategy in areas where the former is common, including areas of sub-Saharan Africa or South Asia<sup>7,17,19</sup>. Because residue retention and crop rotation are generally considered good agronomic practices, a reconsideration of the order in which conservation agriculture principles are introduced in these regions (that is, better targeting of no-till to systems already based on the other two conservation agriculture principles) is not in conflict with general recommendations for sustainable crop production.

Our analysis, synthesizing information from hundreds of field trials across 48 crops and 63 countries (Extended Data Fig. 1), shows that no-till significantly enhances yields (7.3%) under rainfed agriculture in dry climates when the other two conservation agriculture principles are also implemented (Fig. 2a). Yet, the reverse is true when no-till is applied alone (−11.9%). Furthermore, yields decrease with no-till regardless of whether the other principles are applied in humid climates (Fig. 2b). These results are consistent with smaller data sets (for example, 26 studies on no-till rainfed maize) in which residue retention in semiarid environments and crop rotation positively impacted yields<sup>9</sup>. A yield benefit with no-till in combination with the other two conservation agriculture principles in dry climates is probably because of improved water infiltration and greater soil moisture conservation<sup>6,20</sup>. We found that when water is non-limiting owing to irrigation, no-till in dry climates maintains yields similar to conventional systems (residue retention + crop rotation mean effect size for 34 studies and 213 observations: −3.0%; 95% confidence interval: −6.2 to 0.4%), providing further support for this conclusion.

To help meet current and future crop production challenges, our results suggest that this set of integrated management practices can provide agronomic benefits in water-limited and/or water-stressed regions. This is an important finding given that millions of hectares in dry climates of sub-Saharan Africa and South Asia have recently been identified as suitable for sustainable intensification efforts<sup>21</sup>. Still, if conservation agriculture is to be successful at increasing crop productivity in these areas, it must be adjusted to local conditions through an innovative, multi-stakeholder driven approach that is sensitive to market opportunities, equipment availability, and farmers' production objectives and needs<sup>8,17,19</sup>. Our findings further suggest that no-till in combination with the other two conservation agriculture principles, when targeted appropriately, may become an increasingly important strategy to deal with



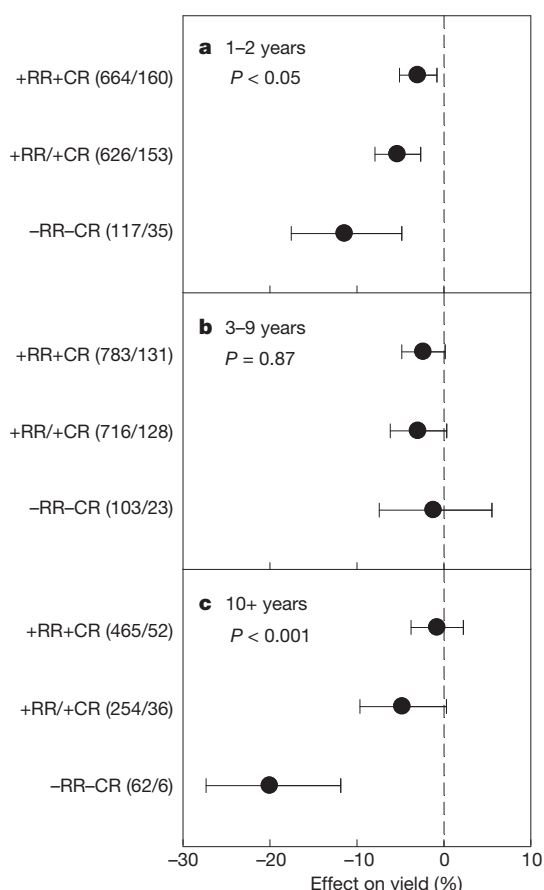
**Figure 2 | Comparison of rainfed crop yield in no-till versus conventional tillage systems in relation to the other two principles of conservation agriculture as a function of climate.**

The influence of (a) 'Dry' and (b) 'Humid' climates, defined by aridity index values (mean annual precipitation divided by potential evapotranspiration) less or more than 0.65, respectively. Categories represent studies that indicated the presence or absence of residue retention and crop rotation for both no-till and conventional tillage systems: +RR+CR (residue retention + crop rotation), +RR/+CR (either residue retention or crop rotation), or -RR-CR (without residue retention or crop rotation). The number of observations and total number of studies included in each category are displayed in parentheses. Error bars represent 95% confidence intervals. Significant differences between categories are indicated by *P* values based on randomization tests.

soil moisture stress due to climate change. Projected changes in precipitation and temperature are expected to cause increased drying and drought in important agricultural production areas of the world<sup>22,23</sup>. Depending on the severity of these changes, a number of adaptations will be required to maintain global agricultural production levels; the attributes of soil moisture conservation and water use efficiency with the three principles of conservation agriculture could play an important role.

It is often suggested that the risk for short-term decreases in crop productivity represents a major barrier for farmers considering conservation agriculture<sup>8–10</sup>. Our results confirm this; regardless of whether the other two conservation agriculture principles are implemented, no-till reduces yields in the first few years following adoption (Fig. 3a). However, the yield decline in initial years is minimized when all three principles are applied compared with one principle (−3.0% versus −11.4%, respectively). Moreover, despite no-till yields in all categories becoming comparable with conventional tillage in the medium term (Fig. 3b), after 10+ years yields begin to decline when only no-till is implemented (Fig. 3c). Hence, to mitigate the negative impacts of no-till, our findings emphasize the importance of implementing all three principles and the overall need for strategies to overcome yield reductions in early and later (10+) years. Although the economic benefits of conservation agriculture may be more strongly driven by cost reductions rather than increased yields<sup>17</sup>, negative yield outcomes can discourage poorer farmers who tend to focus on short-term gains, probably making it an overriding factor limiting the adoption of conservation agriculture<sup>7,19</sup>.

It cannot be determined from our database whether initial yield reductions are caused by biophysical conditions (for example, soil structure, decomposition of residues on the soil surface) or sub-optimal management (that is, a learning curve effect). The transition to no-till integrated with the other two conservation agriculture principles is challenging



**Figure 3 | Comparison of yield in no-till versus conventional tillage systems in relation to the other two principles of conservation agriculture over time.** The influence of (a) 1–2, (b) 3–9, and (c) 10+ years following no-till implementation. Categories represent studies that indicated the presence or absence of residue retention and crop rotation for both no-till and conventional tillage systems: +RR+CR (residue retention + crop rotation), +RR/+CR (either residue retention or crop rotation), or –RR–CR (without residue retention or crop rotation). The number of observations and total number of studies included in each category are displayed in parentheses. Error bars represent 95% confidence intervals. Significant differences between categories are indicated by *P* values based on randomization tests.

as it represents a holistic change in management requiring adaptation at the individual farm-level. A targeted review of no-till studies in sub-Saharan Africa and South Asia reported a high risk of short-term yield declines for major annual crops<sup>10</sup>. Similar to previous work<sup>9</sup>, these authors also noted that implementation of the other two conservation agriculture principles can minimize this risk and that no-till yield losses tend to diminish with time<sup>10</sup>. Interestingly, regardless of initial impacts on yield, our results do not indicate that no-till outperforms conventional tillage in the 10+ year category. One possible explanation for these results is that weed, pest, and disease pressures may increase with continuous no-till systems over time depending on how the other conservation agriculture principles are implemented<sup>24</sup>, possibly offsetting improvements in soil quality. Further research is needed to identify initial and long-term yield constraints of no-till systems.

When considering the relative importance of crop rotation versus residue management practices in enhancing yield of no-till systems, our meta-analysis does not provide evidence that one principle regulates productivity more than the other. Across all observations, the individual effects of residue retention and crop rotation reduce the negative impacts of no-till by 4.8 and 3.8%, respectively, although differences between categories are insignificant (Fig. 1). However, in dry climates these principles each have a much stronger effect on rainfed crop yields, reducing yield losses by 10.1 and 11.0%, respectively. Indeed, previous

work has stressed the importance of residue retention to enhance soil and cropping system benefits of reduced tillage systems<sup>25,26</sup>, with our study being the first to quantify impacts on crop productivity at a global scale. Our results illustrate the need to implement at least one, and preferably both, principles in addition to no-till in rainfed cropping systems in dry climates, while also suggesting that consistent yield declines with no-till in humid environments may be primarily caused by factors unrelated to these principles.

Clearly, there are important environmental (for example, reduced erosion and improved soil quality) and economic outcomes of continuous no-till<sup>15,16,17</sup> beyond the scope of the present analysis that might justify adoption at the farm scale and should be considered in a trade-off analysis against yield reductions. Nevertheless, agricultural regions containing a disproportionate number of the world's poor, including sub-Saharan Africa and South Asia, currently struggle with food security issues and have a high probability of experiencing yield reductions due to climate change in the future<sup>2</sup>. Despite the promising effects of no-till in certain contexts (that is, rainfed agroecosystems in dry climates), we stress that benefits in yield are only seen when the other two conservation agriculture principles are also implemented. Of far greater concern is that no-till alone tends to have the opposite of the intended goal, thereby placing farmers at increased risk of yield losses. It is precisely resource-poor and vulnerable smallholder farming systems that will have the greatest challenges adopting the other two principles, most notably the retention of crop residues due to strong competition for residues by livestock and other uses<sup>8,17</sup>. Hence, efforts to expand conservation agriculture further must remain conscious of the potential for no-till to 'backfire' in these contexts.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 May; accepted 29 August 2014.

Published online 22 October 2014.

- Foley, J. A. *et al.* Solutions for a cultivated planet. *Nature* **478**, 337–342 (2011).
- Lobell, D. B. *et al.* Prioritizing climate change adaptation needs for food security in 2030. *Science* **319**, 607–610 (2008).
- Godfray, H. C. J. & Garnett, T. Food security and sustainable intensification. *Phil. Trans. R. Soc. B* **369**, 20120273 (2014).
- Tilman, D., Balzer, C., Hill, J. & Befort, B. L. Global food demand and the sustainable intensification of agriculture. *Proc. Natl Acad. Sci. USA* **108**, 20260–20264 (2011).
- FAO. *Save and Grow: A Policymaker's Guide to the Sustainable Intensification of Smallholder Crop Production* 1–37 (FAO, 2011).
- Hobbs, P. R., Sayre, K. & Gupta, R. The role of conservation agriculture in sustainable agriculture. *Phil. Trans. R. Soc. B* **363**, 543–555 (2008).
- Stevenson, J. R., Serraj, R. & Cassman, K. G. Evaluating conservation agriculture for small-scale farmers in sub-Saharan Africa and South Asia. *Agric. Ecosyst. Environ.* **187**, 1–10 (2014).
- Giller, K. E., Witter, E., Corbeels, M. & Tittonell, P. Conservation agriculture and smallholder farming in Africa: the heretics' view. *Field Crops Res.* **114**, 23–34 (2009).
- Rusinamhodzi, L. *et al.* A meta-analysis of long-term effects of conservation agriculture on maize grain yield under rain-fed conditions. *Agron. Sust. Dev.* **31**, 657–673 (2011).
- Brouder, S. M. & Gomez-Macpherson, H. The impact of conservation agriculture on smallholder agricultural yields: a scoping review of the evidence. *Agric. Ecosyst. Environ.* **187**, 11–32 (2014).
- Andersson, J. A. & Giller, K. E. in *Contested Agronomy: Agricultural Research in a Changing World* (eds Sumberg, J. & Thompson, J.) Ch. 2, 22–46 (Earthscan, 2012).
- Giller, K. E. *et al.* A research agenda to explore the role of conservation agriculture in African smallholder farming systems. *Field Crops Res.* **124**, 468–472 (2011).
- Friedrich, T., Derpsch, R. & Kassam, A. Overview of the global spread of conservation agriculture. *Field Actions Sci. Rep.* **6**, 1941 (2012).
- Godfray, H. C. *et al.* Food security: the challenge of feeding 9 billion people. *Science* **327**, 812–818 (2010).
- Sachs, J. *et al.* Monitoring the world's agriculture. *Nature* **466**, 558–560 (2010).
- Palm, C., Blanco-Canqui, H., DeClerck, F., Gatere, L. & Grace, P. Conservation agriculture and ecosystem services: An overview. *Agric. Ecosyst. Environ.* **187**, 87–105 (2014).
- Erenstein, O., Sayre, K., Wall, P., Hellin, J. & Dixon, J. Conservation agriculture in maize- and wheat-based systems in the (sub)tropics: lessons from adaptation initiatives in South Asia, Mexico, and Southern Africa. *J. Sustain. Agric.* **36**, 180–206 (2012).

18. Grassini, P. & Cassman, K. G. High-yield maize with large net energy yield and small global warming intensity. *Proc. Natl Acad. Sci. USA* **109**, 1074–1079 (2012).
19. Corbeels, M. *et al.* Understanding the impact and adoption of conservation agriculture in Africa: a multi-scale analysis. *Agric. Ecosyst. Environ.* **187**, 155–170 (2014).
20. Serraj, R. & Siddique, K. H. M. Conservation agriculture in dry areas. *Field Crops Res.* **132**, 1–6 (2012).
21. International Center for Research in the Dry Areas (ICARDA) Geoinformatics Unit. <http://gu.icarda.org/en/> (2014).
22. Cook, B. I., Smerdon, J. E., Seager, R. & Coats, S. Global warming and 21<sup>st</sup> century drying. *Clim. Dyn.* (in the press).
23. Dai, A. Increasing drought under global warming in observations and models. *Nature Clim. Change* **3**, 52–58 (2013).
24. Farooq, M., Flower, K. C., Jabran, K., Wahid, A. & Siddique, K. H. M. Crop yield and weed management in conservation agriculture. *Field Crops Res.* **117**, 172–183 (2011).
25. Govaerts, B. *et al.* Conservation agriculture and soil carbon sequestration: between myth and farmer reality. *Crit. Rev. Plant Sci.* **8**, 97–122 (2009).
26. Erenstein, O. Crop residue mulching in tropical and semi-tropical countries: an evaluation of residue availability and other technological implications. *Soil Tillage Res.* **67**, 115–133 (2002).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We are grateful to the National Key Science and Technology Project of China for supporting X.Q.L. with grant number 2014ZX07101-012.

**Author Contributions** C.v.K., B.A.L., and X.Q.L. conceived the project. All authors contributed to the literature search, except N.v.G. and R.T.V. C.M.P., X.Q.L., J.L., K.J.v.G., B.A.L., and M.E.L. extracted data from publications and contributed to construction of the database. C.M.P., X.Q.L., K.J.v.G., and N.v.G. conducted analyses. C.M.P. and X.Q.L. wrote the manuscript draft and all authors contributed to interpretation of the results and writing of the final paper.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.M.P. (cmpitt@illinois.edu).



## METHODS

**Data collection.** We comprehensively searched the peer-reviewed literature for publications investigating the effects of no-till in relation to the other two conservation agriculture principles on crop yields from Jan 1980 to May 2013 using Scopus (Elsevier). Search terms included 'tillage', 'no till', 'zero till', 'direct drill\*', or 'conservation ag\*' in the article title and 'yield' in the article title, abstract, or keywords. Conference proceedings and non-English language publications were excluded. This search produced a total of 2,471 publications, which were screened on the basis of the following criteria: (1) studies had to represent field experiments containing side-by-side comparisons of no-till and conventional tillage practices; (2) no-till treatments consisted of zero tillage immediately before crop establishment for a given growing season (that is, reduced tillage treatments such as strip-tillage were rejected); (3) crop yields were reported; (4) location of the experiment was stated; (5) management information regarding at least one other conservation agriculture principle was available (that is, residue management or crop rotation); and (6) confounding effects between treatments were absent (that is, differences in residue management, seeding rates, fertilizer rates, etc. were determined to be negligible). When more than one form of tillage was assessed in a study, we selected the treatment representing the greatest soil disturbance (generally mouldboard plow). Although it only represented a small portion of the data, no-till treatments were not always required to represent continuous zero-tillage (for example, if two crops were grown per year and the first crop required tillage but the second was planted using no-till practices, yield comparisons for the second crop were included).

Studies were rejected if it was unclear from reading the experimental methods whether factors other than tillage differed between treatments with the exception of herbicides (the absence of tillage as a weed control strategy generally requires changes in herbicide management under no-till<sup>24</sup>). Owing to the large size of the database, particular attention was given to avoiding data duplication (for example, when different studies reported the same data). Thus, if it was unclear whether a publication contained duplicate data, it was rejected. A number of publications from the conservation agriculture literature were rejected because of the lack of a control treatment that satisfied our criteria (that is, conservation agriculture treatments representing all three principles were compared with conventional tillage treatments with residues removed and no rotation).

Means for no-till and conventional tillage yields were extracted from each study in addition to study and site characteristics including crop type, study location, study duration, irrigation, residue management, and crop rotation practices. In cases where yield data were only presented in figures, values were extracted using Plot Digitizer (<http://plotdigitizer.sourceforge.net/>). In a few instances where yield data were only reported as a percentage change relative to the other treatment, we assumed absolute yield values for the reference treatment and calculated the natural log of the response ratio normally as described below (because this metric only quantifies the relative difference between means, the same value is produced regardless of the absolute magnitude of means). To investigate changes in yield over time, the number of years since the initiation of no-till was recorded for each observation. Observations were excluded from the analysis of no-till duration when only mean yields over a number of years were presented. For the small number of studies in which tillage occurred periodically in no-till systems, the duration of no-till was reset to time zero with each tillage event. For example, if tillage occurred during the rice phase of a rice–wheat rotation during five consecutive years, each wheat yield observation was recorded as the first year under no-till.

The effect of climate was assessed for yields under rainfed conditions by determining the aridity index (mean annual precipitation divided by potential evapotranspiration) for each study using latitude and longitude coordinates and the WorldClim database<sup>27</sup>. Following the generalized climate classification scheme<sup>28</sup>, aridity index values less and more than 0.65 were categorized as 'dry' and 'humid', respectively. If latitude and longitude coordinates were not stated, an attempt was made to contact study authors. Otherwise, coordinates were estimated using the

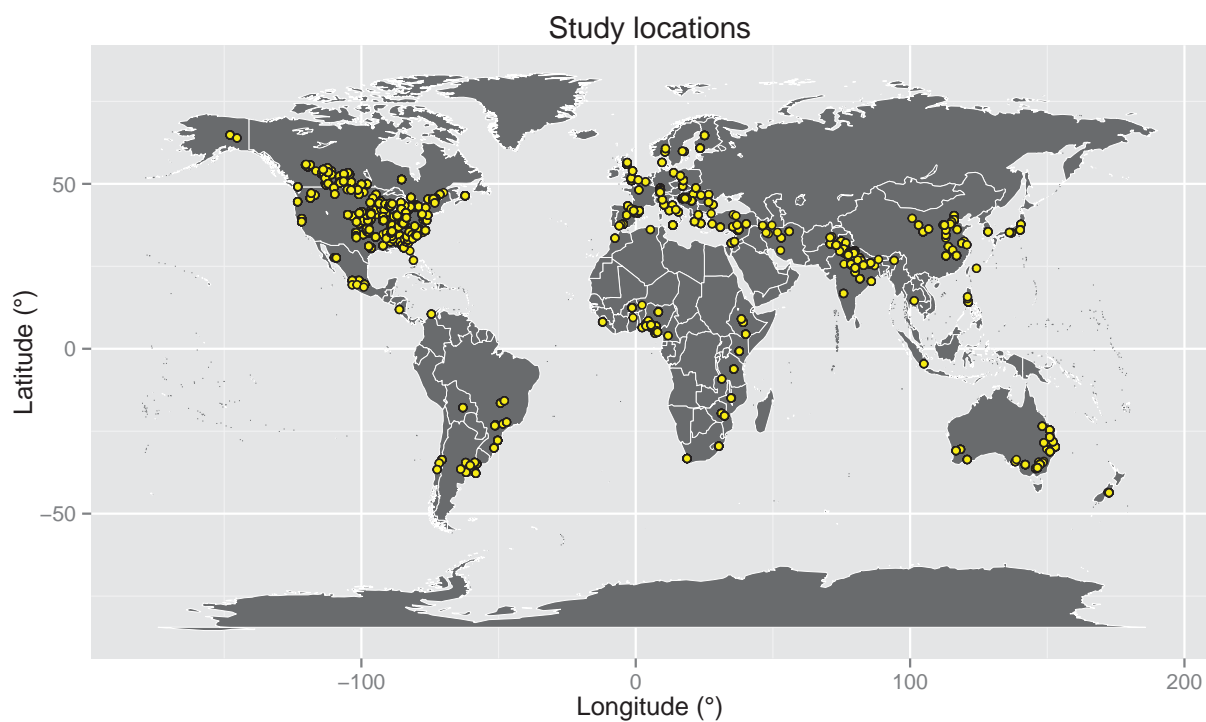
location of the nearest city or the experiment station at which the study took place. In a few cases where only large geographical areas were stated in publications, coordinates and aridity index values were not estimated.

Crop rotation, residue management, and irrigation practices were recorded for each study as categorical variables where possible. Crop rotation was treated as a binary variable (yes/no), where 'yes' indicated that two or more crops were grown in sequence in the same field over time (including the use of a cover crop), and 'no' indicated continuous cultivation of a single crop. Residue management was also treated as a binary variable (retained/removed), where 'retained' indicated that crop residues were retained in the field following harvest each growing season (or in a minority of cases, residues were supplied from elsewhere or by growing a cover crop between seasons), and 'removed' indicated that residues were physically removed from the field or burned following harvest. Information for categorical variables was extracted from the Materials and Methods section of publications, and to a lesser extent was inferred from discussions of crop management details found in the Introduction or Discussion sections. Irrigation practices (yes/no) were recorded when available, with cells left blank when irrigation practices were unclear.

Before data analysis, observations were grouped into categories depending on the presence of residue management and crop rotation practices as determined by information reported in the original studies. For each paired yield observation, we required that no-till and conventional tillage treatments received the same residue management and rotation practices. Thus, categories represented the following comparisons: three principles (no-till versus conventional tillage, both with residues retained + crop rotation), two principles (no-till versus conventional tillage, both with either residues retained or crop rotation), and one principle (no-till versus conventional tillage, both without residue retention or crop rotation).

**Data analysis.** Following previous work<sup>29</sup>, we calculated the natural log of the response ratio (the ratio of no-till to conventional tillage yields) as the effect size in our meta-analysis. Because within-study variance measures for mean yields were available for less than a few percent of studies, individual observations were weighted by replication, with weights =  $(n_{\text{conv.}} \times n_{\text{no-till}}) / (n_{\text{conv.}} + n_{\text{no-till}})$ , where  $n_{\text{conv.}}$  and  $n_{\text{no-till}}$  are the number of replicates for conventional tillage and no-till treatments, respectively<sup>30</sup>. In situations where more than one observation from a study was included in a category, weights were divided by the total number of observations from that study. When yield values for a treatment equalled zero and thereby indicated crop failure or experimental error, observations were excluded. Moreover, observations more than five standard deviations from the weighted mean effect size within each category were excluded (this represented <0.5% of data on average). All statistical analyses were conducted with R (version 3.0.2)<sup>31</sup>. Bootstrapping procedures within the 'boot' package<sup>32</sup> were used to generate 95% confidence intervals for weighted mean effect sizes using 4,999 iterations<sup>30</sup>. Between-group heterogeneity was assessed using randomization procedures based on 4,999 replications<sup>30</sup>. Results were considered significant if confidence intervals did not overlap with zero and randomization tests yielded  $P$  values <0.05. For ease of interpretation, all results were back-transformed and reported as percentage change in yield for no-till relative to conventional tillage practices.

27. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).
28. UNEP. *World Atlas of Desertification*, 2nd edn (eds Middleton, N. & Thomas, D.) (Edward Arnold, 1997).
29. Hedges, L. V., Gurevitch, J. & Curtis, P. S. The meta-analysis of response ratios in experimental ecology. *Ecology* **80**, 1150–1156 (1999).
30. Adams, D. C., Gurevitch, J. & Rosenberg, M. S. Resampling tests for meta-analysis of ecological data. *Ecology* **78**, 1277–1283 (1997).
31. R Core Team. R: A Language and Environment for Statistical Computing. <http://www.R-project.org> (R Foundation for Statistical Computing, 2013).
32. Canty, A. & Ripley, B. boot: Bootstrap R (S-Plus) Functions. R package v.1.3-11 (2014).



Extended Data Figure 1 | The location of studies containing yield comparisons between no-till and conventional tillage systems used in the meta-analysis.

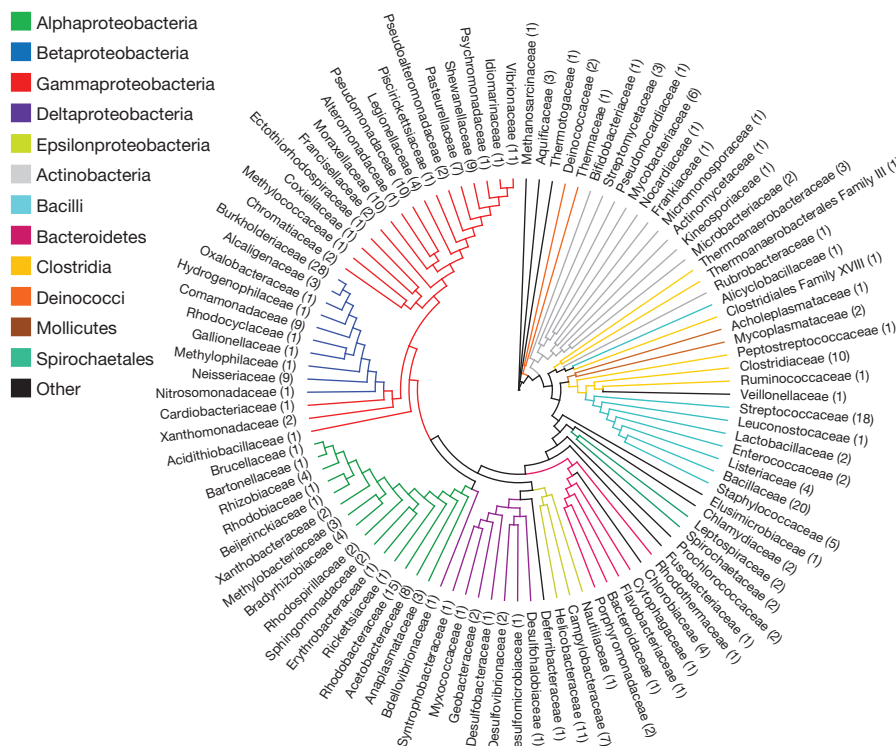
# Long-term phenotypic evolution of bacteria

Germán Plata<sup>1,2</sup>, Christopher S. Henry<sup>3</sup> & Dennis Vitkup<sup>1,4</sup>

For many decades comparative analyses of protein sequences and structures have been used to investigate fundamental principles of molecular evolution<sup>1,2</sup>. In contrast, relatively little is known about the long-term evolution of species' phenotypic and genetic properties. This represents an important gap in our understanding of evolution, as exactly these proprieties play key roles in natural selection and adaptation to diverse environments. Here we perform a comparative analysis of bacterial growth and gene deletion phenotypes using hundreds of genome-scale metabolic models. Overall, bacterial phenotypic evolution can be described by a two-stage process with a rapid initial phenotypic diversification followed by a slow long-term exponential divergence. The observed average divergence trend, with approximately similar fractions of phenotypic properties changing per unit time, continues for billions of years. We experimentally confirm the predicted divergence trend using the phenotypic profiles of 40 diverse bacterial species across more than 60 growth conditions. Our analysis suggests that, at long evolutionary distances, gene essentiality is significantly more conserved than the ability to utilize different nutrients, while synthetic lethality is significantly less conserved. We also find that although a rapid phenotypic evolution

is sometimes observed within the same species, a transition from high to low phenotypic similarity occurs primarily at the genus level.

Analyses of phenotypic evolution, such as the morphological variation of beaks in Darwin's finches<sup>3</sup>, provided the original impetus and context for understanding natural selection. Because the evolutionary importance and physiological role of specific phenotypic traits change over time, it is often difficult to connect genotype to phenotype to fitness across long evolutionary distances, especially for metazoan organisms. For microbial species, on the other hand, the ability to metabolize different nutrient sources, although clearly not the only important phenotype, always remains an essential determinant of their fitness and lifestyle. Even though a large-scale comparative analysis of microbial phenotypes—such as growth on different nutrients or the impact of genetic perturbations—is currently challenging owing to a relative paucity of experimental data, we rationalized that thoroughly validated computational methods can be used to investigate the phenotypic evolution of diverse bacterial species. Flux balance analysis (FBA)<sup>4</sup>, in particular, has been previously used to accurately predict gene and nutrient essentiality, growth yields, and evolutionary adaptations to environmental and genetic perturbations<sup>5</sup>. Notably, the accuracy of FBA methods has



**Figure 1 | Diversity of considered bacterial families.** The cladogram shows the evolutionary relationship between the 100 bacterial families that include the 322 species considered in our study. The tree is based on the average 16S ribosomal RNA (rRNA) genetic distances between species in each family

(see Methods). The numbers of considered species in each family are shown in parentheses. Different colours represent different bacterial classes. The tree was rooted using the *Methanosarcina barkeri* rRNA sequence.

<sup>1</sup>Department of Systems Biology, Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032, USA. <sup>2</sup>Integrated Program in Cellular, Molecular, Structural and Genetic Studies, Columbia University, New York, New York 10032, USA. <sup>3</sup>Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA. <sup>4</sup>Department of Biomedical Informatics, Columbia University, New York, New York 10032, USA.

been independently demonstrated for many dozens of species encompassing diverse phylogenetic distributions and growth environments<sup>6</sup>. We selected for our analysis more than 300 phylogenetically diverse bacteria (Fig. 1) for which genome-scale metabolic models were reconstructed using a recently developed protocol<sup>7</sup> (see Methods).

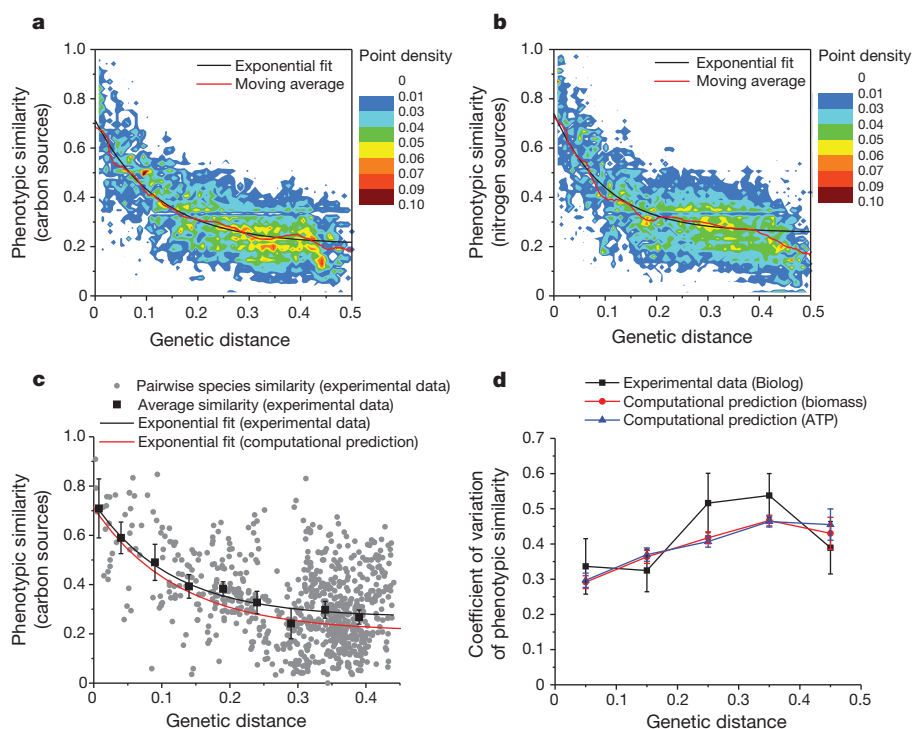
To investigate the long-term evolution of growth phenotypes, we considered 62 carbon sources that are commonly used by microbial species for growth and energy production<sup>8</sup>. For each considered species we used FBA to determine a subset of the compounds that could be used for biomass synthesis or ATP generation—two key metabolic objectives of bacterial growth<sup>9</sup>. This analysis resulted in binary phenotypic profiles that describe the ability of each microbial species to use each of the considered compounds (see Methods). The evolution of these phenotypic profiles—that is, the change in phenotypic similarity as a function of species divergence (genetic distance)—is shown in Fig. 2a, (see also Extended Data Fig. 1). Notably, this analysis demonstrated that the average long-term evolution of growth phenotypes can be approximated well by an exponential decay (see Methods and Extended Data Table 1). A three-parameter exponential model fits the data in Fig. 2a significantly better than simpler alternative models (Extended Data Table 2). Similar divergence trends were observed for larger sets of carbon source compounds (Extended Data Fig. 2), and for compounds that could be used as a source of nitrogen (Fig. 2b). The observed trend was also robust towards subsampling or removal of specific species and families used in the analysis (Extended Data Fig. 3).

The observed exponential trends suggest that as microbial species diverge over planetary timescales and adapt to different environmental niches, approximately similar fractions of phenotypic properties change per unit time. For species separated by more than 1 billion years of

evolution ( $\sim 0.2$  genetic distance in Fig. 2), the divergence of growth phenotypes approaches saturation around a similarity of 21% (Fig. 2a), which is higher than the value expected by chance ( $\sim 12\%$ ) given the average number of carbon compounds used by the models. This difference is likely due to a widespread utilization of common nutrient sources across bacterial species (see Extended Data Table 3)<sup>10</sup>.

Notably, before the evolution of growth phenotypes settles into the aforementioned average trend, a much higher rate of phenotypic evolution is observed for pairs of bacteria at very close genetic distances ( $< 0.01$ , or  $\sim 50$  million years<sup>11</sup>). Our computational analysis predicts  $\sim 71\%$  phenotypic similarity for closely related bacteria (Fig. 2, Extended Data Table 1), which agrees well with available experimental data on intra-species phenotypic similarity: for example, 75% for the utilization of carbon sources in *Escherichia coli*<sup>12</sup> and 69% for *Campylobacter jejuni*<sup>13</sup>. The diversity of bacteria observed at close distances reflects a well-documented genetic and phenotypic variability within bacterial pangenomes<sup>14</sup>. The observed patterns also suggest that phenotypic evolution proceeds in two different stages, namely through fast phenotypic diversification of closely related strains followed by a slower exponential divergence lasting billions of years. Notably, patterns of multi-stage and hierarchical evolution have been observed in other systems, for example in bacterial and eukaryotic developmental networks<sup>15,16</sup>.

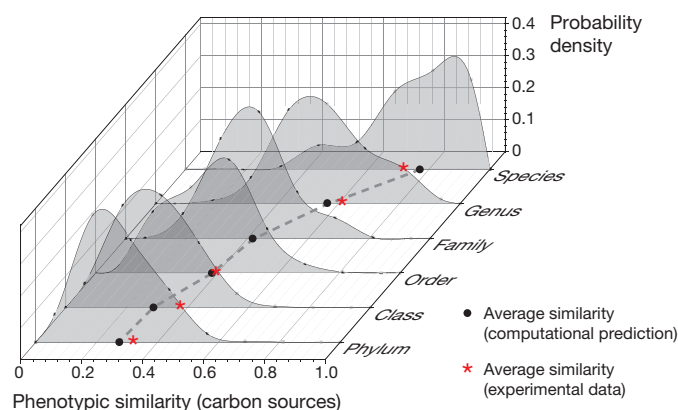
To validate experimentally the predicted patterns of long-term phenotypic evolution of bacteria, we obtained GENIII Biolog Phenotype Microarrays<sup>8</sup> data for 40 diverse microbial species (Extended Data Fig. 4 and Supplementary Data 1). Phenotype Microarrays data were used to determine the ability of each considered bacteria to utilize the 62 different carbon sources used in the simulations (Fig. 2c; see Methods). In agreement with previous results<sup>7</sup>, FBA predicted microbial growth phenotypes



**Figure 2 | Evolution of bacterial metabolic growth phenotypes.** Genetic distances in the figure are based on bacterial 16S rRNA sequences. **a**, The evolution of phenotypic similarity in the usage of carbon sources for biomass synthesis. The colours represent the point density at a given genetic distance for all pairwise comparisons between metabolic models ( $n = 26,106$ ). The black line shows a three-parameter exponential fit to the computational predictions; the red line shows a moving average of the predictions. **b**, Like **a**, but phenotypic similarity in the usage of nitrogen sources for biomass production across metabolic models ( $n = 36,856$ ). **c**, Experimental analysis of the long-term phenotypic divergence trend. Grey points represent pairwise comparisons of

carbon usage phenotypes (Biolog data) between 40 bacterial species ( $n = 780$ ). The black squares represent the average values of experimental phenotypic similarity at different divergence distances. The black line represents an exponential fit to the experimental phenotypic similarity data; the red line represents an exponential fit to the computationally predicted phenotypic similarity data for biomass synthesis. **d**, The variability of experimental and computationally predicted phenotypic similarity at different divergence distances. The variability was quantified by the coefficient of variation, defined as the ratio of the standard deviation to the mean. Error bars in **c** and **d** represent s.e.m. obtained on the basis of 10,000 bootstrap re-samplings of the considered species.





**Figure 3 | Distribution of phenotypic similarity at different levels of bacterial taxonomic classification.** The distributions of phenotypic similarity in the usage of carbon sources for biomass synthesis were obtained based on computational simulations of metabolic models ( $n = 26,106$ ). The dashed line connects the average values (black dots) of computational predictions at each taxonomic level. The red asterisks in the figure indicate the average values of experimental data obtained using Biolog arrays (see Methods).

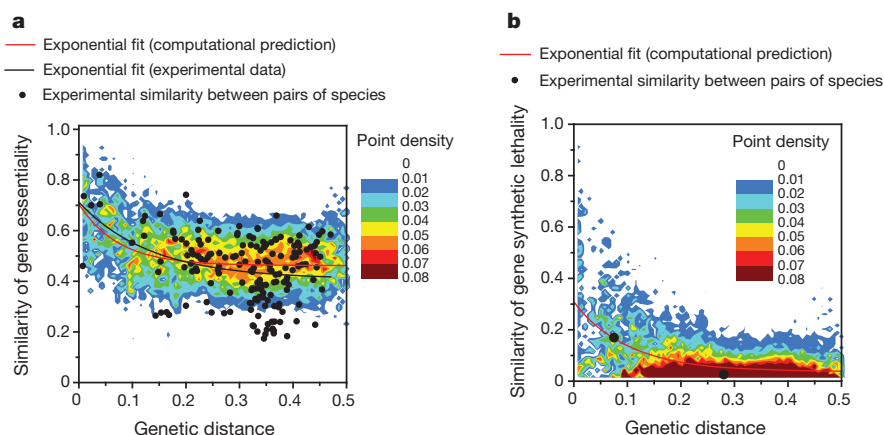
with an average accuracy of  $\sim 70\%$ . Importantly, the experimental results agree well with the computationally predicted average trend describing phenotypic bacterial divergence (Fig. 2c). Also, for the experimental data, as well as for the computational predictions, the three-parameter exponential model fitted the data significantly better than simpler alternative models (Extended Data Table 2). The comparison of computational and experimental values for the coefficient of variation of phenotypic similarity shows that computational predictions capture well not only the average trend but also the variability of phenotypic similarity for bacteria at different genetic distances (Fig. 2d). Overall, the analyses of experimental data suggest that although individual models need to be further validated and improved, high-throughput metabolic reconstructions can be used for comparative functional studies across a large number of diverged species.

We next investigated the diversity of metabolic growth phenotypes at different levels of conventional taxonomic classification (Fig. 3). Although bacteria from the same species show mostly similar phenotypic properties, the long left tail of the top distribution in Fig. 3 suggests that some organisms have substantial phenotypic differences even at this basic taxonomic level. At the genus level the distribution is very broad, with an average similarity of  $\sim 60\%$ ; this suggests that transitions from high

to low phenotypic similarity usually occur at the level of genera. On the contrary, much lower conservation levels are observed for taxonomic ranks beyond the level of families, where the differences between the ranks are relatively small. This analysis suggests that computational approaches similar to the one presented here could be useful in refining bacterial taxonomy.

To complement the analysis of metabolic growth phenotypes, we used FBA to investigate the long-term evolution of gene deletion phenotypes (Fig. 4). Specifically, we considered the evolution of metabolic gene essentiality and synthetic lethality (see Methods). First, we confirmed a high ( $\sim 76\%$ ) accuracy of FBA gene essentiality predictions for considered species with available experimental data<sup>7</sup> (Extended Data Table 4). Second, our analysis demonstrated that the average long-term evolution of gene essentiality can also be approximated by an exponential divergence (Fig. 4a). Notably, the average rate of evolution of metabolic gene essentiality is substantially faster and saturates at closer genetic distances than the evolution of growth phenotypes (Fig. 2, see Extended Data Table 1). Even at long evolutionary distances, for an average pair of microbial species more than half of the conserved essential genes in one species usually remain essential in the other. Reassuringly, the predicted average trend (Fig. 4a, red line) is consistent with available experimental data (Extended Data Table 4) for microbial species with genome-wide gene deletion screens (Fig. 4a, black dots/black line).

In contrast to gene essentiality, our analysis revealed a very low conservation of synthetic lethality between metabolic genes (Fig. 4b). Following a common definition, we considered a pair of non-essential genes to be synthetic lethal if simultaneous *in silico* deletion of the corresponding reactions from FBA models made biomass synthesis infeasible. Even at close genetic distances ( $<0.01$  in Fig. 4b) synthetic lethality is conserved, on average, for only  $\sim 30\%$  of orthologous metabolic gene pairs. At close distances there is also a substantial variability in the conservation of synthetic lethality across species. As bacterial species diverge further, the average conservation of synthetic lethality drops to  $\sim 5\%$ . This suggests that synthetic lethality is much more sensitive to changes in microbial genotypes than gene essentiality and metabolic growth phenotypes. Only several comprehensive studies, none of them in bacteria, have been performed to assess experimentally the conservation of genetic interactions and synthetic lethality<sup>17,18</sup>. Comparison of fitness data from budding and fission yeast revealed a conservation of epistatic gene pairs of  $\sim 29\%$  (ref. 17) (corresponding to  $\sim 17\%$  similarity). On the other hand,  $\sim 5\%$  of the orthologues of synthetic lethal gene pairs in yeast were also found to be synthetic lethal in *Caenorhabditis elegans*<sup>18</sup> ( $\sim 2.5\%$  similarity). Although these data were obtained in eukaryotic species



**Figure 4 | Evolution of bacterial genetic phenotypes.** Genetic distances are based on bacterial 16S rRNA sequences. **a**, The evolution of similarity in gene essentiality across the considered bacterial species. The colours represent the point density at a given genetic distance for pairwise comparisons among considered models ( $n = 48,920$ ). The red line shows a three-parameter exponential fit to the computational predictions. Black points represent the

available gene essentiality experimental data for 21 bacterial species (see Extended Data Table 4) ( $n = 173$ ); the black line shows an exponential fit to the experimental data. **b**, Like **a**, but for the evolution of similarity in synthetic lethality ( $n = 39,616$ ). Black points represent the experimentally determined similarity in synthetic sick/synthetic lethality for two pairs of eukaryotic species (see Methods).

and the FBA accuracy for predictions of genetic interactions is lower than for essentiality or growth phenotypes<sup>19</sup>, the available experimental results (Fig. 4b, black dots) are generally consistent with the average divergence trend predicted in our bacterial simulations.

We note that the observed behaviour of long-term phenotypic divergence is somewhat reminiscent of the molecular clock in protein evolution<sup>1</sup>. Similar to protein evolution, it is likely that the phenotypic divergence trends are due to both bacterial adaptation to diverse environmental niches and neutral changes<sup>20</sup>. The relative contribution of adaptive and neutral changes is likely to be different in each particular lineage and evolutionary context. Our analysis shows that growth phenotypes, gene essentiality, and synthetic lethality diverge with different rates and have different sensitivities to bacterial genotypes. It is likely that many other phenotypic properties, such as the ability to synthesize different compounds, interact with other species, or withstand specific environmental perturbations, will also show distinct evolutionary patterns. We believe that the accelerating pace of genomic and metagenomic sequencing, and continuous improvement in computational annotation methods<sup>21</sup>, will soon allow mapping of the evolution of various phenotypic properties across the entire bacterial tree of life.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 6 May; accepted 2 September 2014.**

**Published online 26 October 2014.**

1. Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins* (eds Bryson, V. & Vogel, H.) 97–166 (Academic, 1965).
2. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
3. Darwin, C. *The Origin of Species* (Barnes & Noble Classics, 2008).
4. Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nature Biotechnol.* **28**, 245–248 (2010).
5. Oberhardt, M. A., Palsson, B. O. & Papin, J. A. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5**, 320 (2009).
6. Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J. & Lee, S. Y. Recent advances in reconstruction and applications of genome-scale metabolic models. *Curr. Opin. Biotechnol.* **24**, 617–623 (2011).
7. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnol.* **28**, 977–982 (2010).
8. Bochner, B. R. Global phenotypic characterization of bacteria. *FEMS Microbiol. Rev.* **33**, 191–205 (2009).
9. Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. & Sauer, U. Multidimensional optimality of microbial metabolism. *Science* **336**, 601–604 (2012).
10. Peregrin-Alvarez, J. M., Sanford, C. & Parkinson, J. The conservation and evolutionary modularity of metabolism. *Genome Biol.* **10**, R63 (2009).
11. Moran, N. A., Munson, M. A., Baumann, P. & Ishikawa, H. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proc. R. Soc. Lond. B* **253**, 167–171 (1993).
12. Sabarwal, V. *et al.* The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity. *J. Evol. Biol.* **24**, 1559–1571 (2011).
13. Gripp, E. *et al.* Closely related *Campylobacter jejuni* strains from different sources reveal a generalist rather than a specialist lifestyle. *BMC Genomics* **12**, 584 (2011).
14. Monk, J. M. *et al.* Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl Acad. Sci. USA* **110**, 20338–20343 (2013).
15. de Hoon, M. J., Eichenberger, P. & Vitkup, D. Hierarchical evolution of the bacterial sporulation network. *Curr. Biol.* **20**, R735–R745 (2010).
16. Kirschner, M. W. & Gerhart, J. C. *The Plausibility of Life: Resolving Darwin's Dilemma* (Yale Univ. Press, 2005).
17. Dixon, S. J. *et al.* Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc. Natl Acad. Sci. USA* **105**, 16653–16658 (2008).
18. Tischler, J., Lehner, B. & Fraser, A. G. Evolutionary plasticity of genetic interaction networks. *Nature Genet.* **40**, 390–391 (2008).
19. Szappanos, B. *et al.* An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genet.* **43**, 656–662 (2011).
20. Barve, A. & Wagner, A. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500**, 203–206 (2013).
21. Plata, G., Fuhrer, T., Hsiao, T. L., Sauer, U. & Vitkup, D. Global probabilistic annotation of metabolic networks enables enzyme discovery. *Nature Chem. Biol.* **8**, 848–854 (2012).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank B. Bochner and Biolog for providing the experimental phenotypic growth data. We also thank members of the Vitkup laboratory for discussions. This work was supported in part by the National Institute of General Medical Sciences GM079759 grant to D.V. and the U54CA121852 grant to Columbia University. The work by C.S.H. was supported by the Department of Energy contract DE-AC02-06CH11357, as part of the SB Knowledgebase.

**Author Contributions** G.P. and D.V. conceived the study and performed the research and data analysis. C.S.H. built the metabolic models. D.V. directed the research. G.P. and D.V. wrote the manuscript. All authors read and edited the manuscript.

**Author Information** The 322 models used in this study are available at <http://vitkuplab.c2b2.columbia.edu/phenotypes>. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.V. (dv2121@columbia.edu).

## METHODS

**Metabolic models.** We obtained 322 genome-scale metabolic models using a recently published protocol for automatic network reconstruction<sup>7</sup>. To minimize possible biases due to computational gap-filling and network auto-completion, we only considered models in which more than 80% of the reactions were directly based on available gene annotations. To prevent biases related to uneven sampling of bacterial phylogenetic space, we did not use models from the order Enterobacteriales, which contains a significantly higher number of sequenced genomes compared with other bacterial lineages. The exact identity of the considered species did not have a significant impact on the predicted average trends (Extended Data Fig. 3).

**Computation of genetic distances.** For the considered bacteria, 16S rRNA gene sequences were obtained from GenBank<sup>22</sup>. Sequences were aligned using Clustal Omega<sup>23</sup> to a reference alignment of small subunit rRNA sequences from the SILVA database<sup>24</sup>. Genetic distances were then calculated on the basis of the multiple sequence alignment using the Dnadist program in the Phylip software package<sup>25</sup>; the F84 model of nucleotide substitution, with default parameter values, was used. The cladograms in Fig. 1 and Extended Data Fig. 4 were computed from the distance matrix using the Fitch program in Phylip with default parameters; the *M. barkeri* rRNA was used as the outgroup sequence. The taxonomic classification used in Fig. 3 was obtained from the National Center for Biotechnology Information taxonomy database<sup>26</sup>.

**Prediction of growth phenotypes.** FBA allows one to determine feasible values of metabolic reaction fluxes subject to reaction stoichiometry constraints and the assumption of metabolic steady state<sup>4</sup>. Additional constraints, such as upper and lower bounds for metabolic fluxes or flux combinations, can be applied to the models. All FBA calculations in our manuscript used the COBRA toolbox<sup>27</sup>. We used the following procedure to determine the ability of each species to utilize the considered nutrient sources. First, we identified a set of compounds required by the models to simulate growth on different carbon or nitrogen sources (Supplementary Data 2); these compounds include various vitamins, nucleotides and amino acids<sup>8</sup> as well as several model-specific requirements (see below). To test the ability of each model to use different carbon sources, an *in silico* growth medium was defined where the aforementioned substances were constrained to a maximum combined uptake of 10 mmol of carbon per gram dry weight, and all other carbon compounds were removed. All carbon-free compounds were made available in the simulated medium (with the maximum uptake rate of 1 mol g<sup>-1</sup> dry weight). An analogous procedure was used to define a growth medium for testing nitrogen sources. Second, for each considered carbon or nitrogen source (compound), we used FBA to calculate the maximum biomass or ATP synthesis rate when the compound was made available in the corresponding *in silico* medium (maximum uptake rate of 1 mol g<sup>-1</sup> dry weight). Similar to the treatment of experimental data (see experimental procedures below), biomass or ATP flux values were normalized on a scale of 0–100 corresponding to the minimum (no carbon or nitrogen source tested) and maximum flux values across considered compounds, respectively. Similar to experimental measurements, carbon and nitrogen sources scoring 10 or above were considered as being positive for growth. The metabolic compounds tested in our analysis correspond to carbon and nitrogen sources that are commonly used by multiple bacteria (Extended Data Table 3) and are assayed in Biolog MicroPlates<sup>8</sup>. To prevent low phenotypic similarities arising because of models with a very low overall number of positive growth phenotypes, we only considered models that could synthesize biomass on more than five of the tested compounds; the exact value of this cutoff had little effect on the observed average trends (Extended Data Fig. 5a). In total, 229 and 272 models were used for the analysis of carbon and nitrogen sources for biomass synthesis, respectively.

**Definition of an *in silico* growth medium.** Similar to bacterial growth *in vivo*, many of the metabolic models used in our analysis are auxotrophic for specific compounds, beyond the main carbon and nitrogen sources tested for their ability to support microbial growth (62 carbon and 68 nitrogen compounds); that is, the models can simulate biomass or ATP synthesis only if small amounts of additional nutrients are available in the simulated growth media. To define a single minimal medium, used across all models to test growth on the main carbon sources, we used the following procedure. First, in addition to the main carbon sources (available with a maximum uptake of 1 mol g<sup>-1</sup> dry weight), all metabolic compounds that could be imported by the models were made available in the simulated media with a maximum combined carbon uptake of 10 mmol g<sup>-1</sup> dry weight; we note that this maximum uptake rate is only 1% of the maximum uptake of the main carbon sources. Second, we determined which of the main carbon sources could support growth under these conditions. Third, for each model the additional carbon compounds were sequentially removed (in a random order) from the simulated media until no compound could be further removed while allowing growth on the main carbon sources determined in the second step. Fourth, the additional carbon sources required for growth in more than 75% of the tests with a positive growth phenotype were combined across all models; this resulted in the carbon-containing

component of the minimal media. Fifth, the same procedure was used to determine the nitrogen-containing component of the minimal media. Sixth, the carbon- and nitrogen-containing components were combined to produce the minimal media used in the study (Supplementary Data 2). Notably, very similar results were observed using different minimal media obtained from independent runs of the aforementioned procedure. Very similar results (Extended Data Fig. 6) were also obtained when the *in silico* growth medium, used for all tests, contained all possible nutrients (without any removals) with maximum uptakes rates of 1 mol g<sup>-1</sup> dry weight for the main carbon or nitrogen sources, and with combined maximum uptake rates of 10 mmol g<sup>-1</sup> dry weight (1% of the uptake for the main nutrients) for additional carbon or nitrogen compounds, respectively.

**Prediction of essentiality and synthetic lethality.** To determine essential genes, we first established the association between every gene and corresponding metabolic reactions. We then simulated gene deletions by setting the maximal fluxes through corresponding reactions that cannot be catalysed by the products of other genes to zero. If such an *in silico* deletion of a gene made it impossible, on the basis of FBA calculations, to produce a non-zero biomass, the gene was considered to be essential. A pair of non-essential genes was considered to be synthetic lethal if simultaneous deletion of the two genes made it impossible to produce a non-zero biomass. All FBA simulations for testing gene essentiality and synthetic lethality were performed, similar to common experimental procedures, using an *in silico* rich medium: that is, non-zero fluxes were allowed through every transport reaction in the models. Genes associated with lumped reactions, such as 'protein synthesis', were not considered in the calculations. To prevent low phenotypic similarities arising because of models with a very small number of essential genes, only models with more than ten predicted essential genes or synthetic lethal gene pairs with mapped orthologues were considered in the analysis; 314 and 290 models were used for the analysis of essentiality and synthetic lethality, respectively.

**Quantifying phenotypic similarity.** For a given set of features, namely carbon sources, essential genes, or synthetic lethal gene pairs, similarities between species were quantified by Jaccard's similarity index. Jaccard's index is defined as the size of the intersection between two given sets divided by the size of the union of the two sets; for example, if A is the set of all carbon sources that can be used by species *a*, and B the set that can be used by species *b*, then Jaccard's carbon source similarity between *a* and *b* is defined as  $J(A,B) = |A \cap B| / |A \cup B|$ . Importantly, to calculate the similarity of gene essentiality and synthetic lethality between species, we only considered orthologous genes and gene pairs that are shared between corresponding metabolic networks. Orthologous genes were identified using bi-directional BLASTP<sup>28</sup> hits (with expect (*E*) values <0.05) between the species' genomes.

**Analysis of experimental Biolog data.** A collection of 40 microbial species spanning a wide range of phylogenetic distances (Extended Data Fig. 4) was used to confirm the computationally predicted trends. The ability of these species to metabolize the 62 carbon sources used in the computational analysis was determined using Biolog GENIII Phenotype Microarrays<sup>8</sup>; all data were obtained directly from the Biolog GEN III database (Biolog). The Phenotype Microarrays technology is based on the reduction of a tetrazolium dye, which allows determination of the usage of different nutrient sources across multiple growth conditions<sup>8</sup>. Biolog assays were performed essentially as described in the GEN III MicroPlate instruction manual. Colorimetric measurements after a 24 h incubation period for each species and each carbon source were expressed on a scale of 0–100, representing the average colour density (across at least five biological replicates) in each well of the Biolog plate relative to the negative and positive controls; only scores of 10 or above were considered as evidence that a tested compound was used by a species. This cutoff value was obtained on the basis of the bimodal-like distribution of the data<sup>12</sup> (Extended Data Fig. 5b); similar results were obtained using other cutoff values (Extended Data Fig. 5c). The accuracy of FBA in predicting microbial growth phenotypes was evaluated using the nine species that were present both in the computational and experimental analyses. The experimental values in Fig. 3 (red stars) were based on data from the aforementioned 40 species, and the intra-species similarity studies in refs 12 and 13.

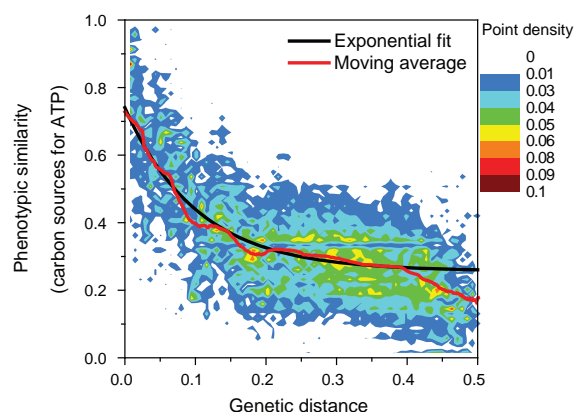
**Experimental gene essentiality data.** Gene essentiality data were compiled for 19 species with genome-wide gene deletion screens (Extended Data Table 4). Species from the genus *Mycoplasma* were excluded from this analysis because of their very small genomes and a very high (~80% (ref. 29)) fraction of essential genes. For every pair of species, orthologous genes were identified using bi-directional BLASTP hits (with *E* values <0.05). The similarity of gene essentiality was determined for genes with identified orthologues annotated as enzymes in the KEGG database<sup>30</sup>. To estimate the similarity of gene essentiality at close genetic distances, we also considered partial essentiality data for *Streptococcus pneumoniae* R6 and *Staphylococcus aureus* N315, which were compared with essential genes in *Streptococcus sanguinis* SK36 and *S. aureus* NCTC 8325, respectively. To use these incomplete data, and given the similarity of the species' genome sizes, we assumed symmetry of essential gene conservation: that is, that the number of essential genes in one species that are

not essential in the other is the same for both bacteria. A similar approach was used to estimate Jaccard's similarity of genetic interactions between eukaryotic species on the basis of published data<sup>17,18</sup> (Fig. 4b).

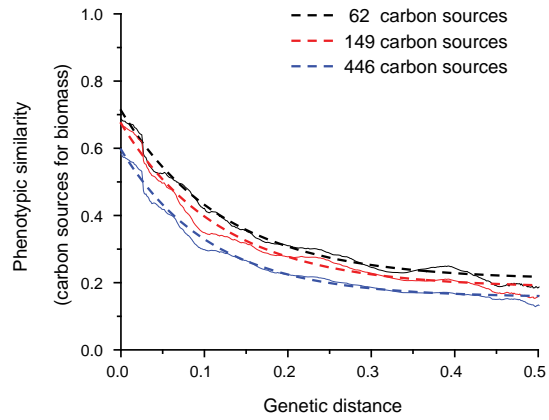
**Exponential model fits.** Pairwise divergence of bacterial phenotypic similarity ( $y$ ) as a function of genetic distance ( $t$ ) was fitted using the following equation:  $y = a + be^{-ct}$ , where the parameter  $a$  represents the saturation level for phenotypic divergence at long genetic distances,  $(a + b)$  represents the level phenotypic similarity at close genetic distances, and the parameter  $c$  quantifies the divergence rate, namely the phenotypic similarity decrease per unit of genetic distance (time). Larger values of  $c$  correspond to faster divergence of the phenotypic similarity. The parameter  $a$  was not considered in the nested two-parameter exponential model used for model comparison. To quantify the genetic distance between bacterial species, we used the divergence between their 16S rRNA sequences; 1% 16S rRNA distance approximately corresponds to 50 million years since divergence from a common ancestor<sup>11</sup>.

22. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **37**, D26–D31 (2009).
23. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
24. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
25. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. *Cladistics* **5**, 164–166 (1989).
26. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37**, D5–D15 (2009).
27. Becker, S. A. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols* **2**, 727–738 (2007).
28. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
29. Glass, J. I. *et al.* Essential genes of a minimal bacterium. *Proc. Natl Acad. Sci. USA* **103**, 425–430 (2006).
30. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

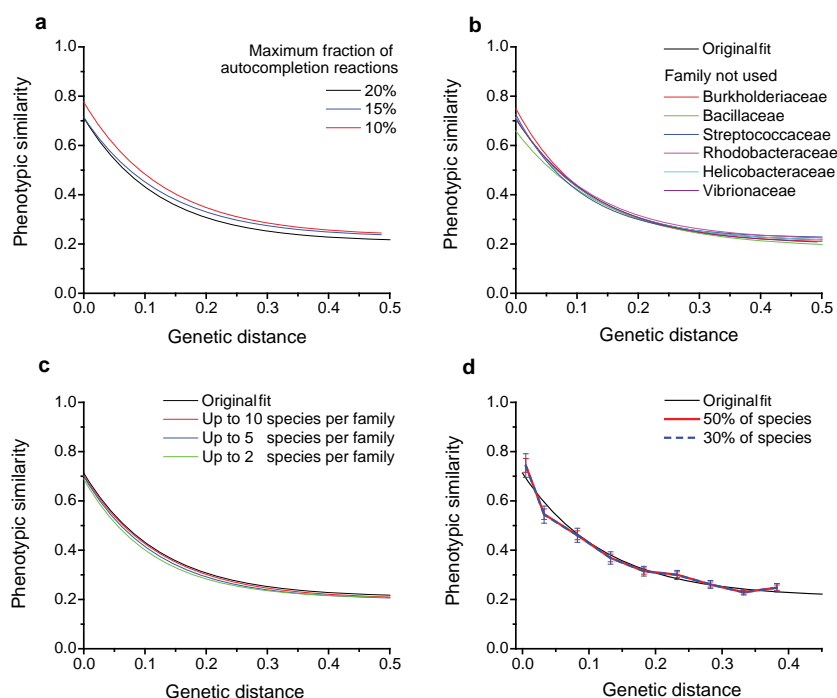




**Extended Data Figure 1 | The evolution of phenotypic similarity in the usage of carbon sources for ATP production.** Genetic distances are based on 16S bacterial rRNA sequences. The colours represent the point density at a given genetic distance for all pairwise comparisons between metabolic models ( $n = 20,910$ ). The black line shows a three-parameter exponential fit to the computational predictions; the red line shows a moving average of the predictions.

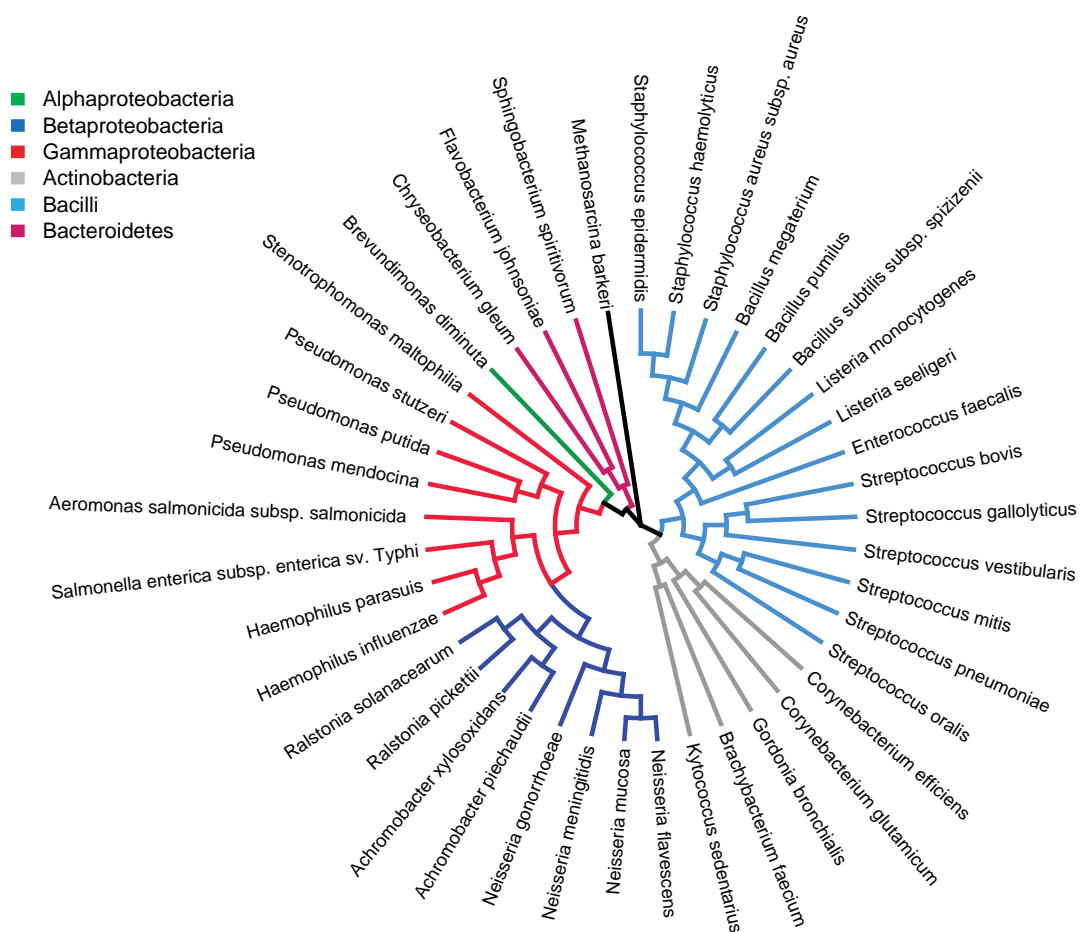


**Extended Data Figure 2 | The evolution of phenotypic similarity for different sets of carbon sources.** Genetic distances are based on 16S bacterial rRNA sequences. Phenotypic similarities of biomass synthesis are shown for 62, 149, and 446 carbon sources. Solid lines represent moving averages (using a 0.05 genetic distance window) of computational predictions; dashed lines represent exponential fits to the data.



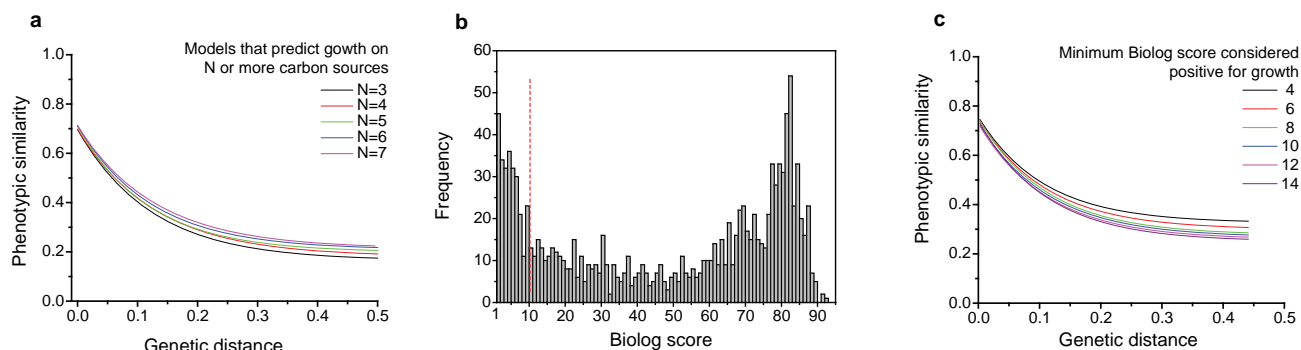
**Extended Data Figure 3 | The effect of species selection on observed patterns of phenotypic divergence.** The black lines in all panels (marked 'Original fit') represent the exponential fit of the phenotypic similarity (carbon source utilization) as a function of genetic distance for all pairs of considered models, that is, models with fewer than 20% auto-completion reactions and more than five predicted carbon sources for growth. The observed trends of phenotypic evolution remain similar when (a) only models with a smaller

fraction of auto-completion reactions are considered, (b) models from individual families that include more than ten modelled species are excluded from the analysis, (c) only a maximum number of species per family is considered, and (d) a subset of species is chosen at random from the pool of all considered models. In d, the average values at different genetic distance bins are shown for 1,000 random samples of a given number of species; error bars represent the s.e.m. obtained on the basis of the 1,000 replicates.



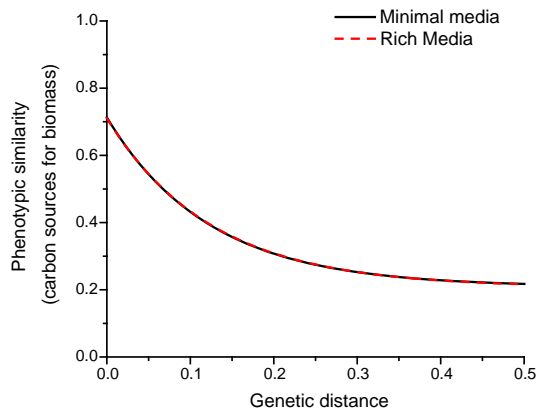
**Extended Data Figure 4 | Evolutionary relationship between 40 species for which experimental phenotype microarray data were considered.** The cladogram is based on 16S bacterial rRNA sequences. Different colours indicate different bacterial classes. The tree was rooted in the *M. barkeri* rRNA sequence.





**Extended Data Figure 5 | The effect of cutoff selection on the computational and experimental phenotypic similarity trends.** **a**, The exponential fits of phenotypic similarity as a function of genetic distance for metabolic models that predict growth on more than a given number of carbon sources. **b**, The frequency distribution of normalized Biolog scores for 40 species across 62 experimental growth conditions. The figure shows a bimodal pattern (scores of

0 are not plotted). The dashed red line shows the cutoff score of 10 used in the main analysis. **c**, The effect of different cutoffs used to define positive growth in the Biolog data. Different lines represent exponential fits to the experimental values of phenotypic similarity based on different values of the Biolog cutoff score.



**Extended Data Figure 6 | Effect of growth media on the predicted phenotypic similarity trends.** The black line shows an exponential fit to the predicted phenotypic similarity in the usage of carbon sources calculated on the basis of the *in silico* minimal media used in the study of growth phenotypes (see Methods and Supplementary Data 2) ( $n = 26,106$ ). The dashed red line shows an exponential fit to the predicted phenotypic similarity using the *in silico* medium in which all carbon sources that could be imported by the models were made available, with the total combined uptake of carbon constrained to a maximum value of  $10 \text{ mM g}^{-1}$  dry weight ( $n = 27,261$ ).

**Extended Data Table 1 | Parameters of the exponential divergence models, describing the evolution of growth and genetic phenotypes**

Phenotype	<i>a</i>	<i>b</i>	<i>c</i>
Carbon source (biomass)	0.21	0.50	8.16
Carbon source (ATP)	0.21	0.50	7.86
Nitrogen source (biomass)	0.25	0.48	9.75
Essentiality	0.46	0.24	13.6
Synthetic lethality	0.037	0.27	9.98

Values in the table show the parameters (*a*, *b*, *c*) of the divergence model  $y = a + be^{-ct}$ , where *y* represents the phenotypic similarity and *t* represents the genetic distance between species. Parameter *a* represents the saturation level of phenotypic divergence at long genetic distances, (*a* + *b*) represents the level phenotypic similarity at close genetic distances, and the parameter *c* quantifies the divergence rate.

**Extended Data Table 2 | Model comparisons for predicted and experimentally determined phenotypic similarity as a function of genetic distance**

Phenotype	3-parameter exponential vs. linear model (relative likelihood based on Akaike weights)	3-parameter exponential vs. 2-parameter exponential model (F-test P-value)
FBA carbon source similarity (biomass)	$>10^{20}$	$< 10^{-20}$
Experimental carbon source similarity (Biolog)	$9 \times 10^9$	$4.6 \times 10^{-8}$
FBA essentiality similarity	$>10^{20}$	$< 10^{-20}$
Experimental essentiality similarity	27.2	$8.5 \times 10^{-3}$
FBA Synthetic lethality similarity	$>10^{20}$	$< 10^{-20}$

Comparisons between the three parameter exponential model and the linear model were performed on the basis of Akaike's Information Criterion (AIC). Values in the table represent the Akaike-based relative likelihoods of the three-parameter exponential model compared with the linear model. Comparisons between the three-parameter exponential model and the nested two-parameter exponential model were performed using the *F*-test; the corresponding *P* values reflect the probability that the nested two-parameter model fits the data as well as the more complex three-parameter model.



**Extended Data Table 3 | The predicted frequency of carbon and nitrogen source usage across metabolic models**

Carbon sources for biomass production		Nitrogen sources for biomass production	
Metabolite name	Number of models	Metabolite name	Number of models
L-Glutamic Acid	208	Ammonia	241
a-D-Glucose	199	Urea	241
D-Fructose	185	L-Proline	226
L-Malic Acid	165	L-Glutamic Acid	203
L-Lactic Acid	164	L-Valine	203
Maltose	149	L-Isoleucine	183
Glycerol	134	L-Leucine	178
L-Serine	128	Nitrate	155
L-Aspartic Acid	126	Nitrite	145
D-Mannose	115	L-Glutamine	136
L-Arginine	112	L-Serine	127
N-Acetyl-DGlucosamine	104	L-Aspartic Acid	124
D-Trehalose	97	Cytosine	123
Sucrose	89	L-Arginine	119
Inosine	87	L-Ornithine	115
a-Keto-GlutaricAcid	84	Uracil	111
L-Histidine	82	N-Acetyl-D-Glucosamine	103
L-Alanine	77	Cytidine	101
D-GlucuronicAcid	74	Adenosine	94
D-Serine	72	Glycine	93
a-D-Lactose	71	L-Methionine	92
Formic Acid	69	Xanthine	87
Acetoacetic Acid	68	Histamine	86
D-Cellobiose	67	L-Histidine	83
D-Mannitol	64	L-Tryptophan	80
D-Malic Acid	64	L-Alanine	79
D-Sorbitol	61	Ethanolamine	74
D-GalacturonicAcid	56	D-Alanine	73
Mucic Acid	55	D-Serine	72
D-Saccharic Acid	55	L-Lysine	61
D-Gluconic Acid	53	D-Glucosamine	61
D-Galactose	47	Putrescine	54
Citric Acid	33	Acetamide	54
D-Raffinose	29	L-Phenylalanine	49
Dextrin	27	Allantoin	47
g-Amino-ButyricAcid	27	L-Tyrosine	45
L-Rhamnose	26	Formamide	44
Salicin	23	Inosine	41
N-Acetyl-b-D-Mannosamine	20	L-Cysteine	25
Glucose-6-Phosphate	20	L-Asparagine	25
D-Melibiose	15	Thymidine	23
D-Aspartic Acid	15	Uridine	23
L-Fucose	13	Guanine	19
M-Inositol	11	Guanosine	18
Quinic Acid	11	N-Acetyl-D-Mannosamine	18
Propionic Acid	11	Xanthosine	16
Stachyose	9	D-Aspartic Acid	15
N-Acetyl-Neuraminic Acid	7	Methylamine	10
D-Arabitol	7	Thymine	9
Acetic Acid	7	L-Threonine	6
Fructose-6-Phosphate	5	D-Lysine	5
N-Acetyl-DGalactosamine	4	Adenine	4
a-Keto-ButyricAcid	4	N-Acetyl-D-Galactosamine	4
Gentiobiose	0	L-Citrulline	3
b-Methyl-D-Glucoside	0	D-Galactosamine	3
D-Fucose	0	D-Glutamic Acid	2
Gelatin	0	Tyramine	2
L-PyroglutamicAcid	0	b-Phenylethylamine	2
Pectin	0	L-Homoserine	1
L-GalactonicAcid-g-Lactone	0	Uric Acid	1
P-HydroxyPhenyl AceticAcid	0	Agmatine	0
a-HydroxyButyric Acid	0	Hydroxylamine	0
		Ethylamine	0
		L-Pyroglutamic Acid	0
		D-Asparagine	0
		D-Mannosamine	0
		D-Valine	0
		Biuret	0

Numbers in the table represent the total number of models, out of 322, predicted to use the corresponding carbon or nitrogen source. Metabolites are ranked from most to least frequent across models.

Extended Data Table 4 | Bacteria with experimental genome-wide data used to analyse the conservation of gene essentiality

Species name	Number of essential genes	Reference PubMed ID
<i>Acinetobacter baylyi</i> ADP1	499	18319726
<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	276	14602916, 12682299
<i>Bacteroides thetaiotaomicron</i> VPI-5482	325	19748469
<i>Burkholderia thailandensis</i> E264	406	23382856
<i>Caulobacter crescentus</i> NA1000	480	21878915
<i>Escherichia coli</i> K-12	302	16738554
<i>Francisella novicida</i> U112	396	17215359
<i>Haemophilus influenzae</i> Rd KW20	667	11805338
<i>Helicobacter pylori</i> 26695	336	15547264
<i>Mycobacterium tuberculosis</i> H37Rv	689	23028335
<i>Porphyromonas gingivalis</i> ATCC 33277	463	23114059
<i>Pseudomonas aeruginosa</i> PAO1	774	14617778
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhi str. Ty2	331	23470992
<i>Salmonella enterica</i> serovar Typhimurium str. SL1344	355	23470992
<i>Shewanella oneidensis</i> MR-1	403	22125499
<i>Sphingomonas wittichii</i> RW1	572	23601288
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> N315	168*	11952893
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> NCTC 8325	351	19570206
<i>Streptococcus pneumoniae</i>	134†	15995353
<i>Streptococcus sanguinis</i>	218	22355642
<i>Vibrio cholerae</i> O1 biovar El Tor str. N16961	344	23901011

\* Only incomplete data available, used to estimate conservation relative to *S. aureus* NCTC 8325.

† Only incomplete data available, used to estimate conservation relative to *S. sanguinis*.

# The neural representation of taste quality at the periphery

Robert P. J. Barretto<sup>1</sup>, Sarah Gillis-Smith<sup>1</sup>, Jayaram Chandrashekar<sup>2</sup>, David A. Yarmolinsky<sup>1</sup>, Mark J. Schnitzer<sup>3</sup>, Nicholas J. P. Ryba<sup>4</sup> & Charles S. Zuker<sup>1,2</sup>

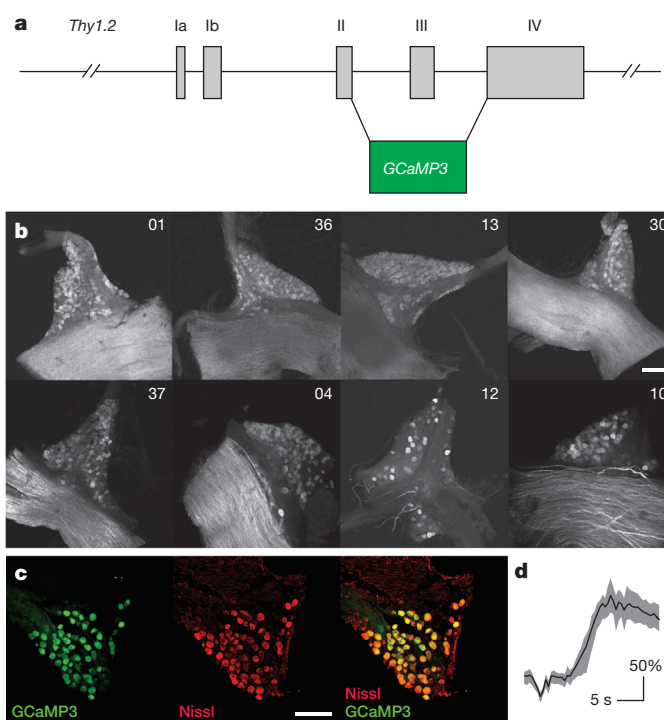
**The mammalian taste system is responsible for sensing and responding to the five basic taste qualities: sweet, sour, bitter, salty and umami. Previously, we showed that each taste is detected by dedicated taste receptor cells (TRCs) on the tongue and palate epithelium<sup>1</sup>. To understand how TRCs transmit information to higher neural centres, we examined the tuning properties of large ensembles of neurons in the first neural station of the gustatory system. Here, we generated and characterized a collection of transgenic mice expressing a genetically encoded calcium indicator<sup>2</sup> in central and peripheral neurons, and used a gradient refractive index microendoscope<sup>3</sup> combined with high-resolution two-photon microscopy to image taste responses from ganglion neurons buried deep at the base of the brain. Our results reveal fine selectivity in the taste preference of ganglion neurons; demonstrate a strong match between TRCs in the tongue and the principal neural afferents relaying taste information to the brain; and expose the highly specific transfer of taste information between taste cells and the central nervous system.**

In mammals, taste receptor cells are assembled into taste buds that are distributed in different papillae in the tongue epithelium. Taste buds are innervated by afferent fibres that transmit information to the primary taste cortex through synapses in the brainstem and thalamus<sup>4</sup>. In the simplest model of taste coding at the periphery, each quality, encoded by a unique population of TRCs expressing specific receptors (for example, sweet cells, bitter cells, and so on), would connect to a matching set of ganglion neurons. Notably, although TRCs are tuned to preferred taste qualities<sup>5–8</sup>, the nature of their functional ‘handshake’ with the nervous system has been a matter of significant debate<sup>1,4,9,10</sup>.

We reasoned that this fundamental question could now be resolved by directly examining the tuning properties of taste ganglion neurons. We focused on the geniculate ganglion, as its neurons innervate all taste buds in the front of the tongue and palate<sup>1</sup>, and opted to use two-photon calcium imaging to monitor tastant-evoked neural activity *in vivo*. This strategy, however, required the solution of two technical challenges: first, the ganglion is located in a bony capsule under the brain, some 4 mm from the surface, far beyond the reach of conventional microscopy; and second, geniculate ganglion neurons would have to be loaded with sensors of neuronal activity that can report function with good temporal, spatial and dynamic range. To solve the first challenge, we implemented the use of two-photon microendoscopy, where a gradient refractive index (GRIN) lens is used as an optical extension device<sup>3</sup>. The GCaMP family of genetically encoded calcium sensors are an attractive tool to solve the challenge of indicator loading<sup>2</sup>, yet there were no suitable mouse lines or drivers appropriate for targeting geniculate ganglion neurons. Therefore, we generated a collection of 40 mouse lines expressing GCaMP3<sup>2</sup> driven by *Thy1* (Fig. 1a), a neuronal promoter highly sensitive to position effects<sup>11</sup>, and screened for those that express the sensor in most geniculate ganglion neurons (Fig. 1b). Line 1 had essentially complete labelling of geniculate ganglion neurons (Fig. 1c), and stimulation of the ganglia *ex vivo* produced reliable calcium-dependent

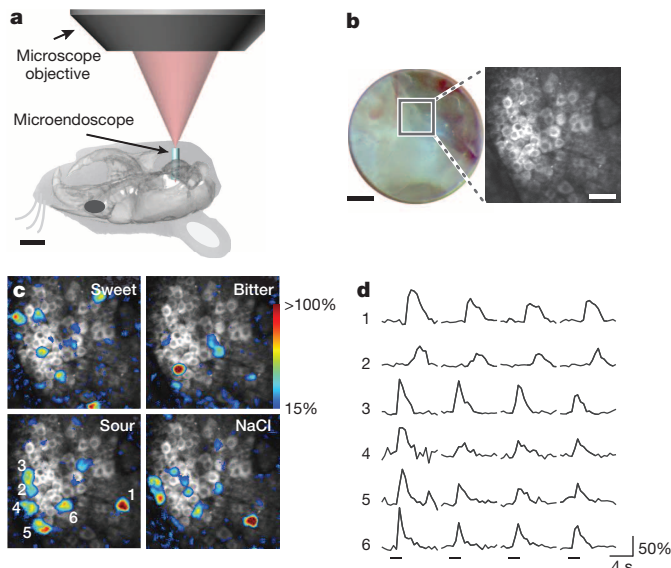
fluorescence changes (Fig. 1d). To take advantage of the more recent versions of GCaMP<sup>12,13</sup>, we subsequently developed a viral infection approach that efficiently labels geniculate ganglion neurons via retrograde transfer of the virus from their terminal fields in the nucleus of the solitary tract (see Methods for details).

To visualize geniculate ganglion neurons in live mice, we exposed a small ventral window into the ganglion (Fig. 2a), and carefully positioned a GRIN lens (1 mm diameter × 3.8 mm length) directly onto the tissue (Fig. 2b). This two-photon imaging configuration allowed unencumbered access to the entire ganglion, and enabled the investigation of the geniculate ganglion at sufficient numerical aperture (~0.45 NA) so as to detect GCaMP-dependent fluorescence changes (Fig. 2c, d).



**Figure 1 | *Thy1*-GCaMP3 transgenic mice express functional GCaMP3 in taste ganglion cells.** **a**, Structure of the *Thy1.2*-GCaMP3 construct<sup>11</sup>. **b**, Whole-mount confocal images of geniculate ganglion from eight transgenic lines (top right) shows GCaMP3 expression in varying subsets of neurons. **c**, Line 1, used in our studies, expresses GCaMP3 in nearly all neurons (>90%,  $n = 6$ ); compare Nissl staining (red) versus GCaMP3 fluorescence (green). **d**, *Ex vivo* calcium imaging of a geniculate from line 1 illustrating strong GCaMP3 responses to a test depolarizing solution (KCl, 500 mM); over 75% of imaged neurons responded with  $\Delta F/F$  greater than 100% ( $n = 25$  cells, mean  $\pm$  quartiles). Scale bars, 100  $\mu\text{m}$ .

<sup>1</sup>Howard Hughes Medical Institute and Departments of Biochemistry and Molecular Biophysics and of Neuroscience, Columbia College of Physicians and Surgeons, Columbia University, New York 10032, USA. <sup>2</sup>Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia 20147, USA. <sup>3</sup>James H. Clark Center, Stanford University, Stanford, California 94305, USA. <sup>4</sup>National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, Maryland 20892, USA.



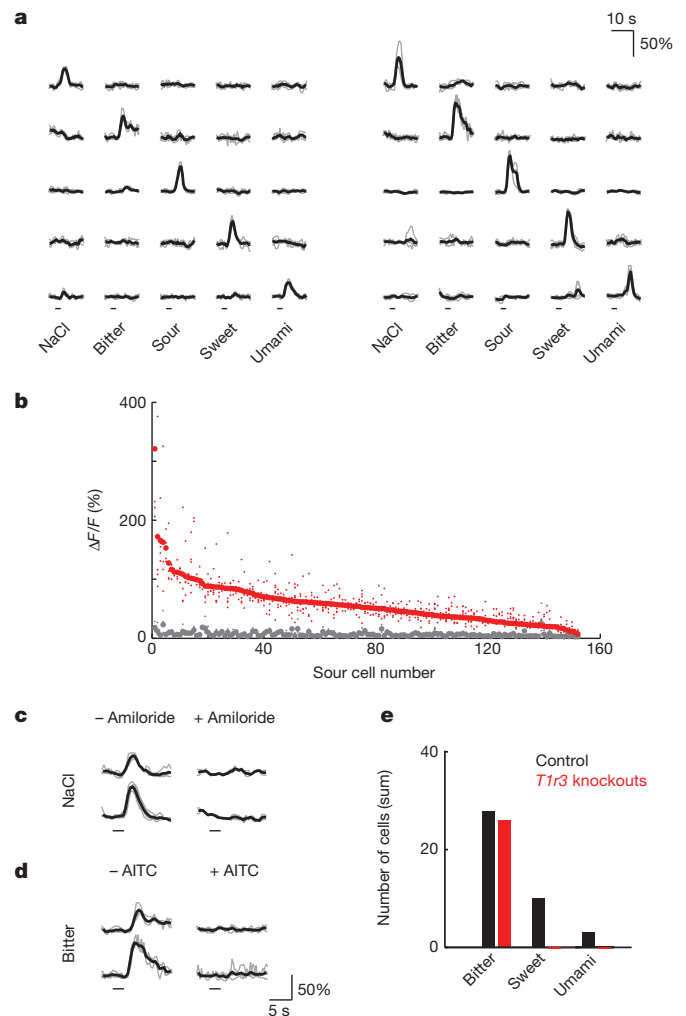
**Figure 2 | In vivo two-photon microendoscopy of the geniculate ganglion.** **a**, Diagram illustrating optical access to the geniculate ganglion. A 1-mm GRIN microendoscope<sup>3</sup> was guided into the surgical opening, and imaged using two-photon microscopy. **b**, Bright-field image through the microendoscope (left), showing individual GCaMP-labelled neurons (right). **c**, Images of a ganglion with 25 GCaMP3 labelled neurons responding to sweet (acesulfame K, 30 mM), bitter (quinine, 5 mM), sour (citric acid, 50 mM) and salt (NaCl, 60 mM) tastants. Fluorescence amplitudes were pseudo-coloured according to  $\Delta F/F$  (scale at right). **d**, Traces from six separate neurons (numbered in **c**) illustrating the time course of amplitude changes in GCaMP3 signals after sour stimulation. Horizontal bars mark the time and duration of the stimulus (inter-stimulus interval was 8 s). Scale bars: **a**, 4 mm; **b**, 200  $\mu$ m and 50  $\mu$ m (magnification).

We assessed the responses of geniculate ganglion neurons to tastants using a range of stimulus paradigms that included the five basic taste qualities at concentrations that evoke strong behavioural and nerve responses<sup>8</sup> (Extended Data Fig. 1). For most recordings the tongue was exposed to a 6.5-s pre-stimulus application of artificial saliva, a 2-s exposure to a test tastant, and a 6.5-s artificial saliva post-stimulus wash. Each session included a minimum of four trials per tastant, and a neuron was classified as a responder if it exhibited statistically significant responses in at least 50% of the trials (see Methods and Extended Data Table 1 for details).

In the tongue, sweet and umami tastes are mediated by a small family of three G-protein-coupled receptors (GPCRs) that combine to form two heteromeric receptors: T1R1 and T1R3 for umami<sup>8,14,15</sup> and T1R2 and T1R3 for sweet<sup>8,14</sup>. Bitters are detected by a family of approximately three-dozen GPCRs, the T2Rs<sup>16</sup>, and sour and sodium taste are sensed by ion channel receptors<sup>5,6</sup>. Given a palette of five different tastes, and all possible neuronal tuning combinations, there are a total of 31 potential neuronal classes: for example, five of these would be tuned to single taste qualities (one for each of the five basic tastes), ten doubly tuned, ten classes responding to three tastes, five to four tastes and a single class tuned to all five tastes. This raises the question of how taste is represented in the ganglion.

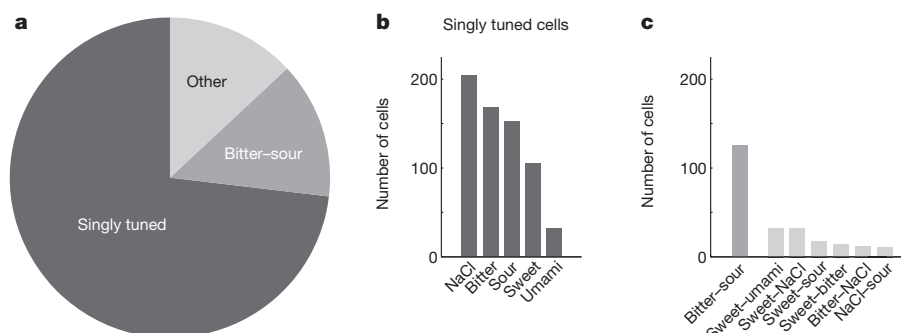
We focused on a standard set of stimuli representing the five basic qualities: sweet-responding neurons were identified using sucrose, bitter with either quinine or cycloheximide, salt by stimulating with NaCl, sour with citric acid, and umami with monopotassium glutamate plus inosine monophosphate. Tastant stimulation of the tongue elicited strong calcium transients in subsets of ganglion neurons (Fig. 2c, d). The responses were robust, reliable (Extended Data Fig. 2) and specific (Fig. 3 and Extended Data Fig. 3). For example, salt responses showed the expected blockage by amiloride<sup>5</sup> (Fig. 3c and Extended Data Fig. 3e), bitter responses were blocked by allyl isothiocyanate (AITC)<sup>17</sup> (Fig. 3d and

Extended Data Fig. 3f), sweet/umami responses were abolished in knock-outs of the T1R3 receptor subunit<sup>8</sup> (Fig. 3e), and acid responses displayed the appropriate sensitivity to PKD2L1-cell silencing<sup>17</sup> (Extended Data Fig. 4). We examined the reproducibility of the taste-evoked responses by measuring variability across trials and, on average, cells responded in at least 75% of the trials (Extended Data Fig. 2). We analysed nearly 1,000 neurons exhibiting tastant-evoked activity: 443 of these were derived from 15 *Thy1-GCaMP3* mice and 432 from 14 AAV-GCaMP6 animals (see Extended Data Table 1); we also included approximately 50 cells each from AAV-GCaMP3 and AAV-GCaMP5 pilot experiments. Our data demonstrate that the vast majority of the responding neurons are strongly activated by only one taste quality, and thus exhibit highly



**Figure 3 | Selective responses of geniculate ganglion neurons.** **a**, Responses of ten ganglion neurons to the five basic taste qualities. Black traces mark average  $\Delta F/F$  from four individual trials per tastant (grey traces); horizontal bars mark the time and duration of the stimulus. **b**, Individual ganglion neurons exhibit strong taste preferences. The graph shows a rank-ordered plot of calcium transient amplitudes for 152 sour-responsive neurons. For each cell, the mean sour response amplitudes (red) and the mean amplitude of its next-strongest tastant response (grey) are shown; minor dots indicate individual trial amplitudes. The vast majority of these sour cells are strongly tuned to sour taste versus any other taste quality. **c**, **d**, Responses are highly selective. The ENaC inhibitor amiloride<sup>5</sup> blocks NaCl responses (**c**), while the bitter TRC inactivator AITC<sup>17</sup> abolishes bitter responses (**d**); individual traces (grey) and average traces (black) are shown for two representative ganglion neurons before (–) and after (+) pharmacological application of the blockers to the tongue. **e**, Mice lacking T1R3 lack ganglion responses to sweet and umami stimuli. Black bars denote control animals ( $n = 8$ ) and red bars *T1r3* knockouts ( $n = 9$ ).





**Figure 4 | Representation of taste in the primary sensory ganglia.**

**a**, Response profile of 904 ganglion neurons from 37 mice according to taste preference; singly tuned cells (dark grey; see **b** for breakdown), bitter-sour cells (grey) and broadly tuned cells (light grey; see Extended Data Table 2). Importantly, the bitter-sour class actually reflects the activity of

T2R-expressing (bitter) TRCs (see text for details). We note that acid responses from bitter cells are not readily visible in whole-nerve recordings<sup>6,17</sup>, probably reflecting sensitivity differences between single-cell imaging and 'bulk' extracellular recordings. **b**, **c**, Distribution of ganglion neurons according to tastant selectivity.

preferred and narrowly tuned tastant selectivity (Figs 3 and 4). Furthermore, the representation of the various taste classes within the ganglion appears random, with no obvious topography or clustering (Extended Data Fig. 5).

The finding that most ganglion neurons are tuned to single taste qualities substantiates a simple match between TRCs and ganglion neurons, with information from the tongue propagated primarily along labelled lines. This is inconsistent with models proposing a general across-fibre model of coding<sup>4</sup>, or that receptor TRCs first signal their information to broadly tuned presynaptic cells, which would then activate ganglion neurons<sup>9,18</sup>. We find no evidence or support for either of these models, as they would require that most of the ganglion neurons respond across multiple taste qualities.

A key prediction of our findings is the expectation that responses to taste mixes should largely reflect the sum of the responses to the individual tastants (that is, as dedicated lines operating independently of each other). Therefore, we tested the representation of binary mixes relative to the responses to the individual components. We chose three pair sets including sweet, bitter, sour and salt tastes and examined their responses. Our results (Extended Data Fig. 6) demonstrate that taste mixes indeed behave like the simple addition of the responses to the individual tastants.

This study also illustrates and uncovers three additional aspects to the representation of taste in the first neural station. First, it reveals that although there are 26 possible combinations of 'multi-tuned' responses (10 doubles, 10 triples, 5 quads and 1 that may potentially respond to all five classes of taste), only small numbers of bona fide multi-tuned neurons are found at over 1% (see Extended Data Table 2). Second, half of the apparently multi-tuned ganglion neurons belong to just one class, namely bitter-sour. And third, umami-responding neurons are divided between umami-alone and umami-sweet responders.

We next questioned how these multi-tuned ganglion neurons might arise. They may be the result of signalling from two or more classes of TRCs converging into a single target neuron, or they may receive inputs from just one TRC type, but appear multi-tuned due to the nature of the stimulus. Indeed, it has long been known that some tastants, although single molecular species, may activate more than one class of TRCs. This is well exemplified by saccharin which powerfully activates sweet TRCs at low concentrations, but begins to activate bitter TRCs at high concentrations (thought to be the basis for the bitter aftertaste of saccharin); another recently reported example is the activation of bitter cells by potassium salts (KCl) at high concentrations<sup>17</sup>. Thus, we wondered whether the activation of bitter-sour ganglion neurons by acid (representing by far the largest class of apparently multi-tuned neurons; Fig. 4c) may also reflect tastant cross-talk, for instance a subset of T2R receptors in bitter cells being sensitive to low pH, as has been observed for a number of GPCRs<sup>19</sup>.

We reasoned that by using a combination of genetic manipulation and pharmacology it should be possible to explain the origin of these

responses. For example, if bitter-sour neurons receive convergent input from T2R and PKD2L1 cells, then ablating or silencing the PKD2L1-expressing sour TRCs should eliminate the bitter-sour doubly tuned class (that is, by eliminating their sour responses). In contrast, if these neurons reflect bitter cells which are sensitive to acidic stimuli, such genetic manipulation should have no impact on the activity or numbers of bitter-sour-responding ganglion neurons. We imaged mice in which we silenced synaptic transmission from PKD2L1-expressing cells by genetically targeting tetanus toxin<sup>17</sup> (PKD2L1-tetanus). As expected the sour-alone neuronal responses were abolished, but, notably, the bitter-sour class was preserved (Extended Data Fig. 4). These results demonstrate that bitter-sour ganglion neurons do not receive input from sour-sensing TRCs, and strongly argue that the bitter-sour neurons are the result of T2R-expressing bitter TRCs being sensitive to acid stimuli. If true, then blocking bitter taste responses should now eliminate the bitter-sour class. As predicted, Extended Data Fig. 4 shows this to be the case, further demonstrating that the vast majority of all ganglion neurons do receive input from single classes of TRCs.

A surprise that emerged from this study was the finding that there were very few umami ganglion neurons, with many also activated by sweet tastants (Fig. 4). Sweet and umami taste receptors are formed by the combination of three GPCRs from the same gene family<sup>8,14,15</sup>. Interestingly, unlike humans, who can readily distinguish between sweet and umami tastes, rodents exhibit considerable behavioural cross-generalization between these two appetitive taste qualities<sup>20</sup>, even though they still represent sweet and umami in two distinct but closely apposed cortical fields<sup>21</sup>. We suggest that this cross-generalization results in part from convergence of sweet and umami inputs into common ganglion neurons, perhaps as a consequence of their shared origin, and receptor repertoire. Future studies examining the fate, tuning and connectivity of individual TRCs throughout their life cycle should help resolve this question.

The capacity to examine the activity of large ensembles of neurons with single-cell selectivity, even in a tissue buried deep in the head, has afforded a comprehensive examination of the behaviour of gustatory neurons in coding responses to the basic taste qualities with exquisite resolution. The development of this imaging preparation has addressed a long-standing question in the taste field, namely the direct demonstration of labelled lines between TRCs and the brain. It should also provide a powerful experimental platform to identify molecular markers defining each class of taste ganglion neurons, help dissect the nature of their 'handshake' with the tongue and the brainstem, and explore the constancy of taste perception in light of the remarkably rapid turnover of TRCs<sup>22</sup>.

Finally, we note that ganglion neurons not only convey taste quality, but also valence. In this regard, it would be of interest to determine if some of the neurons may be dedicated to carry attractive and aversive signals (that is, valence) rather than tastant 'identity' and thereby delineate the

circuit that mediates immediate taste ingestion or rejection responses, independent of conscious perception and the cortex<sup>23</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 14 May; accepted 19 September 2014.**

**Published online 5 November 2014.**

1. Yarmolinsky, D. A., Zuker, C. S. & Ryba, N. J. P. Common sense about taste: from mammals to insects. *Cell* **139**, 234–244 (2009).
2. Tian, L. *et al.* Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nature Methods* **6**, 875–881 (2009).
3. Jung, J. C., Mehta, A. D., Aksay, E., Stepnoski, R. & Schnitzer, M. J. *In vivo* mammalian brain imaging using one- and two-photon fluorescence microendoscopy. *J. Neurophysiol.* **92**, 3121–3133 (2004).
4. Simon, S. A., de Araujo, I. E., Gutierrez, R. & Nicolelis, M. A. The neural mechanisms of gustation: a distributed processing code. *Nature Rev. Neurosci.* **7**, 890–901 (2006).
5. Chandrashekar, J. *et al.* The cells and peripheral representation of sodium taste in mice. *Nature* **464**, 297–301 (2010).
6. Huang, A. L. *et al.* The cells and logic for mammalian sour taste detection. *Nature* **442**, 934–938 (2006).
7. Mueller, K. L. *et al.* The receptors and coding logic for bitter taste. *Nature* **434**, 225–229 (2005).
8. Zhao, G. Q. *et al.* The receptors for mammalian sweet and umami taste. *Cell* **115**, 255–266 (2003).
9. Chaudhari, N. & Roper, S. D. The cell biology of taste. *J. Cell Biol.* **190**, 285–296 (2010).
10. Frank, M. & Pfaffmann, C. Taste nerve fibers: a random distribution of sensitivities to four tastes. *Science* **164**, 1183–1185 (1969).
11. Feng, G. *et al.* Imaging neuronal subsets in transgenic mice expressing multiple spectral variants of GFP. *Neuron* **28**, 41–51 (2000).
12. Akerboom, J. *et al.* Optimization of a GCaMP calcium indicator for neural activity imaging. *J. Neurosci.* **32**, 13819–13840 (2012).
13. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
14. Li, X. *et al.* Human receptors for sweet and umami taste. *Proc. Natl Acad. Sci. USA* **99**, 4692–4696 (2002).
15. Nelson, G. *et al.* An amino-acid taste receptor. *Nature* **416**, 199–202 (2002).
16. Adler, E. *et al.* A novel family of mammalian taste receptors. *Cell* **100**, 693–702 (2000).
17. Oka, Y., Butnaru, M., von Buchholtz, L., Ryba, N. J. P. & Zuker, C. S. High salt recruits aversive taste pathways. *Nature* **494**, 472–475 (2013).
18. Tomchik, S. M., Berg, S., Kim, J. W., Chaudhari, N. & Roper, S. D. Breadth of tuning and taste coding in mammalian taste buds. *J. Neurosci.* **27**, 10840–10848 (2007).
19. Ghanouni, P. *et al.* The effect of pH on  $\beta_2$  adrenoceptor function. Evidence for protonation-dependent activation. *J. Biol. Chem.* **275**, 3121–3127 (2000).
20. Heyer, B. R., Taylor-Burds, C. C., Tran, L. H. & Delay, E. R. Monosodium glutamate and sweet taste: generalization of conditioned taste aversion between glutamate and sweet stimuli in rats. *Chem. Senses* **28**, 631–641 (2003).
21. Chen, X., Gabbito, M., Peng, Y., Ryba, N. J. P. & Zuker, C. S. A gustotopic map of taste qualities in the mammalian brain. *Science* **333**, 1262–1266 (2011).
22. Beidler, L. M. & Smallman, R. L. Renewal of cells within taste buds. *J. Cell Biol.* **27**, 263–272 (1965).
23. Grill, H. J. & Norgren, R. The taste reactivity test. II. Mimetic responses to gustatory stimuli in chronic thalamic and chronic decerebrate rats. *Brain Res.* **143**, 281–297 (1978).

**Acknowledgements** We thank the National Institute of Dental and Craniofacial Research (NIDCR) transgenic-core and C. Guo at Janelia Farms for help in generating the *Thy1-GCaMP3* mouse lines, B. Shields for histology support, and Y. Oka and M. Butnaru for nerve recording and pharmacological advice. We also thank members of the Zuker laboratory for helpful comments. This research was supported in part by the intramural research program of NIDCR (N.J.P.R.). C.S.Z. is an investigator of the Howard Hughes Medical Institute and a Senior Fellow at Janelia Farms.

**Author Contributions** R.P.J.B. designed the study, carried out the imaging experiments, analysed data and wrote the paper; S.G.-S. developed viral gene delivery to ganglion neurons and characterized the transgenic lines; J.C. characterized the transgenic lines, carried out initial imaging experiments and analysed data; D.A.Y. collected and analysed data; M.J.S. provided microendoscopy expertise; N.J.P.R. and C.S.Z. designed the study, analysed data and wrote the paper.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.S.Z. (cz2195@columbia.edu) or N.J.P.R. (nick.ryba@nih.gov).

## METHODS

**Transgenic animals and mouse strains.** All procedures were carried out in accordance with the US National Institutes of Health (NIH) guidelines for the care and use of laboratory animals, and were approved by the Columbia University, Janelia Farm Research Campus, or National Institute of Dental and Craniofacial Research Animal Care and Use Committees. Reported data were obtained from mice ranging from 2–12 months of age and from both genders; randomization and blinding methods were not used. To generate *Thy1-GCaMP3* mice, complementary DNA encoding GCaMP3 was cloned into the XhoI site of a standard *Thy1* vector<sup>11</sup>. More than 40 lines were generated by pronuclear injection with about 25 expressing significant GCaMP3 in subsets of neurons in the cranial ganglia and/or brain (Fig. 1). Calcium-imaging experiments reported here used line 1, which robustly expresses GCaMP3 in all geniculate ganglion neurons. All other mouse strains have been described previously<sup>8,17</sup>. Sample sizes were chosen to allow robust statistical analysis of data, no statistical method was used to determine sample size.

**Confocal imaging of isolated geniculate ganglia.** Mice were perfused and geniculate ganglia were exposed by removing the brain and bone dorsal to the ganglia as described previously<sup>24</sup>. The greater superficial petrosal and facial nerves were cut, and the ganglia were extracted by blunt dissection. To quantify the number of Thy1-GCaMP3-positive neurons per ganglia, fluorescent Nissl staining (NeuroTrace 530/615, Molecular Probes) was used to label neurons, and the ratio of Nissl-positive to GCaMP3-positive neurons determined; greater than 90% of all neurons are labelled by Thy1-GCaMP3 ( $n = 6$  ganglia). Ganglia were mounted in Vectashield mounting medium and imaged with a Zeiss 510 confocal microscope (Zeiss  $\times 10$ , 0.45 NA microscope objective). For control KCl stimulations, dissected ganglia were submerged in imaging buffer (Hank's Buffered Salt Solution with  $\text{Ca}^{2+}$  and 10 mM HEPES). Suture thread was used to mount the ganglia onto a coverslip within a custom imaging chamber. A depolarizing solution (potassium chloride, 500 mM) was applied to cells between washes with imaging buffer.

**Nerve recordings.** Recording procedures and analysis were carried out as described previously<sup>8,15</sup>.

**Viral delivery of GCaMP sensors.** An alternative strategy to express GCaMP in geniculate ganglion neurons involved injection of AAV constructs in the terminal field of taste-responsive neurons in the brainstem. Mice were anaesthetized with ketamine and xylazine ( $100 \text{ mg kg}^{-1}$  and  $10 \text{ mg kg}^{-1}$ , intraperitoneal), with subsequent booster doses to maintain depth of anaesthesia. Body temperature was controlled using a closed-loop heating system. A small craniotomy ( $< 1 \text{ mm}$  diameter) centred approximately 6.5 mm dorsal to the bregma and 1.25 mm from the midline was performed. AAV carrying GCaMP constructs (AAV9-hSyn-GCaMP3, AAV1-hSyn-GCaMP6s and AAV9-hSyn-GCaMP5G; Penn Vector Core) was delivered to the brain at three locations along the rostrocaudal axis: 1.25 mm/–4.0 mm (lateral coordinate relative to bregma/inferior coordinate relative to the dura), and –6.3 mm, –6.5 mm or –6.7 mm (anterior coordinates relative to bregma); approximately 200 nl was delivered per injection. After incision closure and recovery from surgery, mice were housed in their home cages for at least two weeks before imaging.

**Surgical preparation for *in vivo* imaging.** A metal bar was affixed to the dorsal cranium of an anaesthetized mouse (see above) with cyanoacrylate as described previously<sup>25</sup>. The mouse was positioned in a supine position, and its head rigidly secured using the metal bar. A tracheotomy was performed to maintain a clear airway during tastant delivery to the oral cavity<sup>15</sup>. The surgical strategy used to image the geniculate ganglion *in vivo* was as previously described for rat<sup>26</sup>.

***In vivo* microendoscopy.** A singlet microendoscope probe<sup>25</sup> (1 mm diameter, 0.42 pitch, 700  $\mu\text{m}$  working distance in water) was used to access the deep tissue. To mount the probe stably over the imaging area, melted agarose (1.5–2.0% low-melting-point agarose) was applied to the exposed cavity. The microendoscope probe was carefully positioned over the geniculate ganglia with forceps, using a stereoscope for visual guidance. The probe was held mechanically with forceps until secured by the hardened agarose. A conventional two-photon microscope (Prairie Technologies) using a Ti:Sapphire laser tuned to 920 nm was used for fluorescence excitation. A long-working-distance objective (Olympus  $\times 20$ , 0.4 NA) was used to couple light into the microendoscope. Images were typically acquired at 2–4 Hz (10–30  $\mu\text{s}$  pixel dwell time) over a  $\sim 175$ –350  $\mu\text{m}$  field of view, unless otherwise noted. We did not utilize multiple optical planes to avoid double-counting of cells.

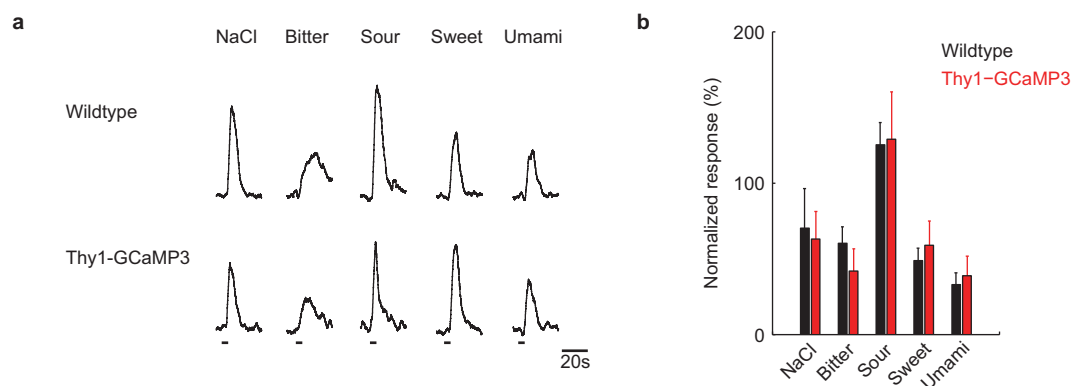
**Tastant delivery.** Lingual stimulation and recording procedures were performed as previously described<sup>8,15</sup>. Tastants were delivered (5–12.5 ml per min) using a feeding

tube positioned approximately 8 mm inside the oral cavity, dorsal to the tongue. Tastants dissolved in artificial saliva were delivered in serial order, interspersed with pure artificial saliva. Recordings were performed during epochs of continuous irrigation; we did not detect calcium activity in response to artificial saliva application, nor did ganglion neurons exhibit mechanical responses. The concentrations of tastants used were: acesulfame K, 30 mM; sucrose, 300 mM; cycloheximide 100–1000  $\mu\text{M}$ ; quinine, 5 mM; sodium chloride, 60 mM; citric acid, 50 mM; monopotassium glutamate + inosine monophosphate, 50 mM + 1 mM, respectively. Pharmacological compounds were delivered as previously described<sup>5,17</sup>: amiloride, 0.010 mM, 8 ml delivered over 1 min; and allyl isothiocyanate (AITC), 3 mM, 40 ml delivered over 5 min followed by artificial saliva rinse for 1 min.

**Calcium imaging data analysis. Image processing.** Two-photon images were first motion-corrected by using cross-correlation-based image alignment<sup>27</sup> (Turboreg, ImageJ plugin). Responding neurons were identified with a semi-automated script that uses independent components analysis (ICA) to examine the spatial and temporal skewness of pixels within active cells<sup>28</sup>. After cell segmentation, all putative responding cells were verified by visually examining the raw imaging data. Cells not detected by ICA analysis were manually segmented. For some cellular regions of interest, there was spatial overlap with a neighbouring cell; if this was minor, they were manually segmented, otherwise they were discarded from further analysis.

**Scoring of tastant-responsive cells.** We used two independent methods to identify tastant-evoked transients in ganglion cells. First, we visually scored cells, by directly observing the aligned image data displayed as a relative fluorescence movie, as well as the putative cells' fluorescence time series. As a control, we used an automated technique to identify transients in the same cells that were visually classified. The data shown in the manuscript used the manual method, although both were largely interchangeable (manual:automatic: sweet cells, 105:95; bitter cells, 168:170; sour cells, 152:152; salt cells, 204:204; umami, 32:26). Because bona fide calcium transients may significantly bias the estimates of location by mean and of dispersion by standard deviation, we used robust statistical measures: the median and  $Q_n$  estimator<sup>29</sup> (median absolute deviation also provides similar results). The  $Q_n$  estimator was calculated for each trial to quantify baseline fluorescence levels. The distribution of  $Q_n$  estimator values was used to estimate the typical relative fluorescence background levels of multiphoton microendoscopy in the geniculate ganglion to be  $0.04 \pm 0.01$ . We visually examined the raw images in outlier trials where background levels exceeded 10%  $\Delta F/F$  (2.5 times  $Q_n$ ). Such trials typically contained inaccurate image registration and were excluded from further analysis. We assumed transients to be outliers of an underlying normal distribution of calcium levels during a cell's resting state. Thus, transients were operationally defined as continuous time intervals where the fluorescence intensities significantly deviated from baseline levels (for example, modified Z-score  $> 2$ ) for at least one fluorescence decay constant (for example, 650 ms for GCaMP3). In both manual and automated cases, we defined transients as tastant-evoked if their onsets occurred between the start of tastant delivery and the two seconds after the end of tastant delivery. Transient onsets varied across animals and are believed to be due to factors associated with tastant arrival: placement of the tube within the oral cavity, tastant flow rate, and location of TRCs relative to the flow. Cells were classified as tastant-responsive if they responded in at least 50% of presentations for a given tastant. Finally, these cells were grouped by their responses to the five basic qualities (any of 31 potential categories, see text). Any category in which the number of contained cells was less than 1% of the total population was excluded from analysis (Fig. 4).

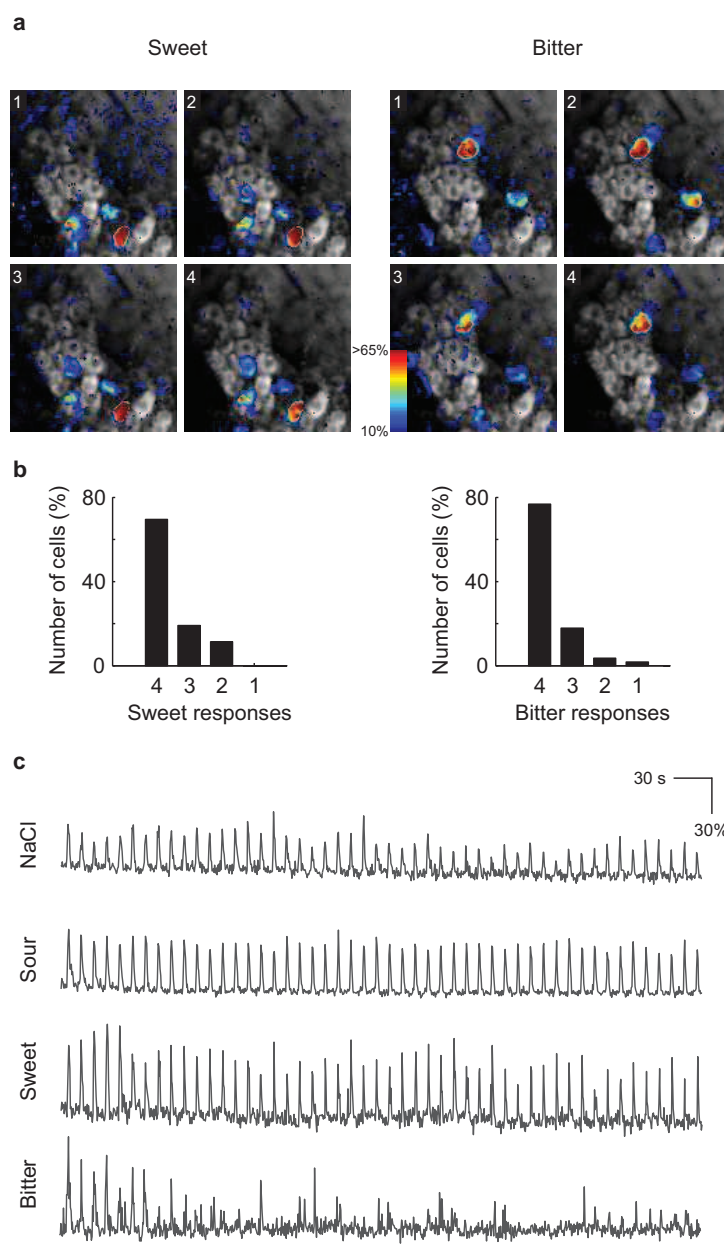
24. Zaidi, F. N. & Whitehead, M. C. Discrete innervation of murine taste buds by peripheral taste neurons. *J. Neurosci.* **26**, 8243–8253 (2006).
25. Barretto, R. P. *et al.* Time-lapse imaging of disease progression in deep brain areas using fluorescence microendoscopy. *Nature Med.* **17**, 223–228 (2011).
26. Sollars, S. I. & Hill, D. L. *In vivo* recordings from rat geniculate ganglia: taste response properties of individual greater superficial petrosal and chorda tympani neurones. *J. Physiol. (Lond.)* **564**, 877–893 (2005).
27. Thévenaz, P., Ruttimann, U. E. & Unser, M. A pyramid approach to subpixel registration based on intensity. *IEEE Trans. Image Process.* **7**, 27–41 (1998).
28. Mukamel, E. A., Nimmerjahn, A. & Schnitzer, M. J. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron* **63**, 747–760 (2009).
29. Rousseeuw, P. J. & Croux, C. Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.* **88**, 1273–1283 (1993).
30. Hyman, A. M. & Frank, M. E. Effects of binary taste stimuli on the neural activity of the hamster chorda tympani. *J. Gen. Physiol.* **76**, 125–142 (1980).



**Extended Data Figure 1 | Thy1-GCaMP3 mice show normal physiological responses to tastants.** **a**, Representative nerve recording traces from control and *Thy1-GCaMP3* mice in response to various tastants (see Methods for details). **b**, Quantification of neural responses (mean + s.e.m.) show that *Thy1-GCaMP3* mice ( $n = 4$ ) are indistinguishable from wild-type mice

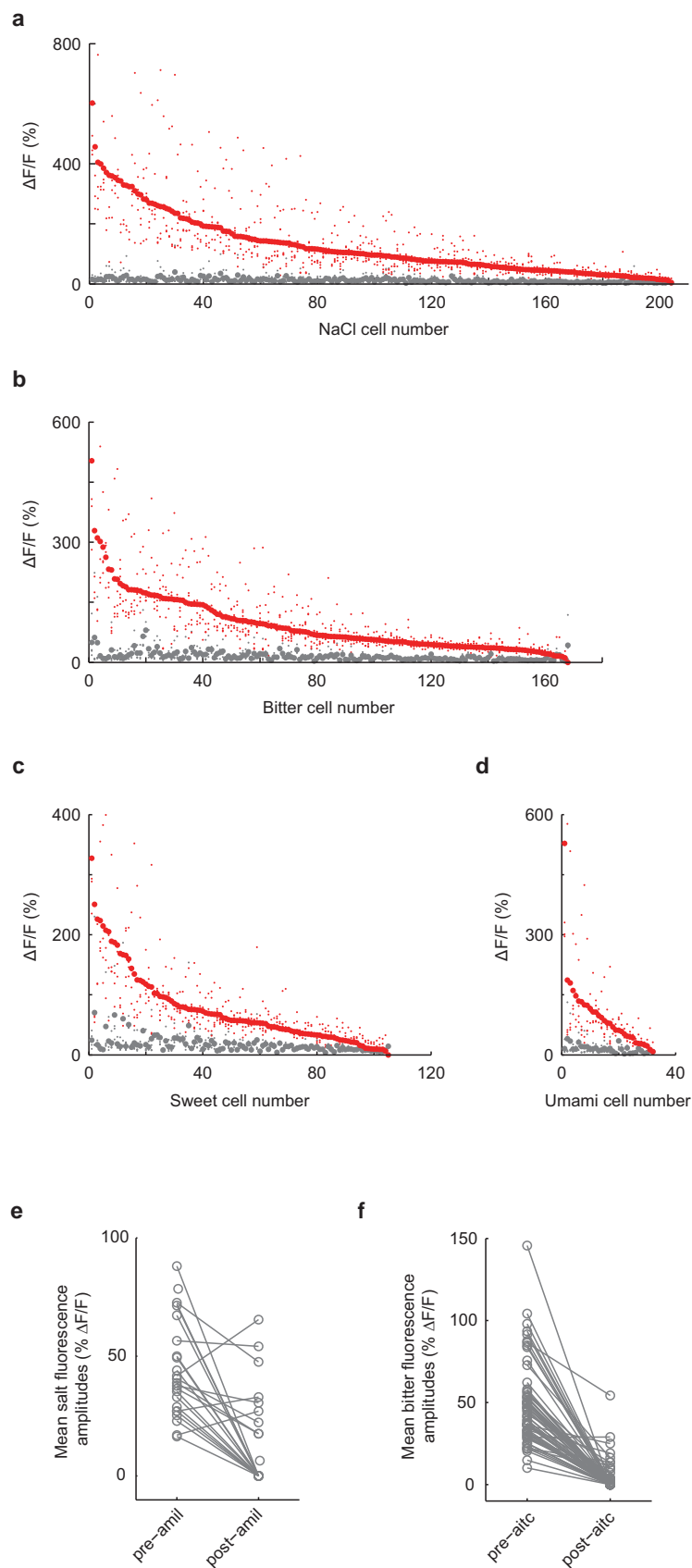
( $n = 3$ ; Student's  $t$ -test; NaCl,  $P = 0.85$ ; bitter,  $P = 0.46$ ; sour,  $P = 0.94$ ; sweet,  $P = 0.69$ ; umami,  $P = 0.77$ ). Recordings were normalized to responses to KCl (500 mM). Horizontal bars below the traces mark the time and duration of the stimulus.





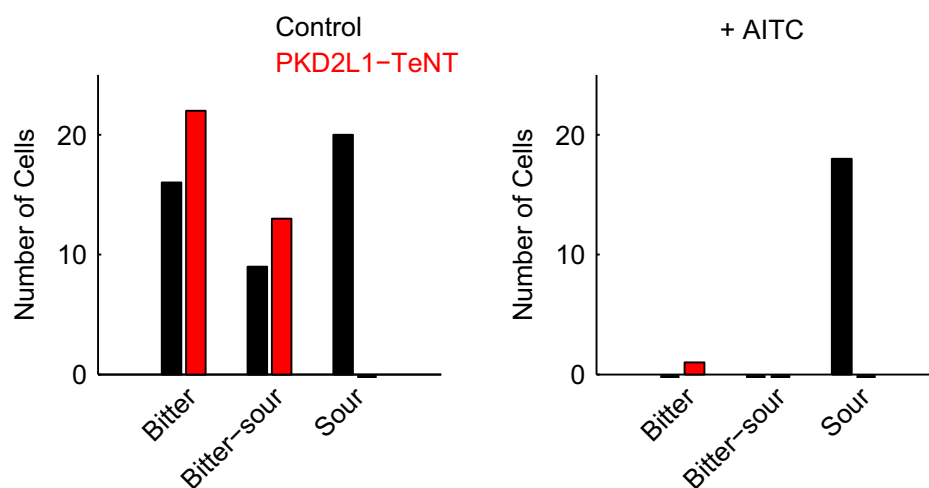
**Extended Data Figure 2 | Reproducibility of tastant-evoked responses in geniculate ganglion neurons.** **a**, Representative images of calcium-evoked GCaMP3 activity in response to sweet (left) and bitter (right) stimulation. Four relative fluorescence images are shown from separate trials. In each trial, the identical cell populations were activated. **b**, We tested 105 sweet responding cells and 168 bitter responding cells for their reproducibility in our automated scoring algorithm for four trials. The histograms show the number of times the

cells respond all four times, three out of four, two out of four, and one out of four. **c**, Sample traces of four representative neurons challenged with 50 trials of the same tastant over a time window of 10 min. Note the high reliability in the activation of the neurons. This experiment also illustrates the desensitization of bitter neurons (bottom traces) over time. Horizontal bars below the traces mark the time and duration of the stimulus.



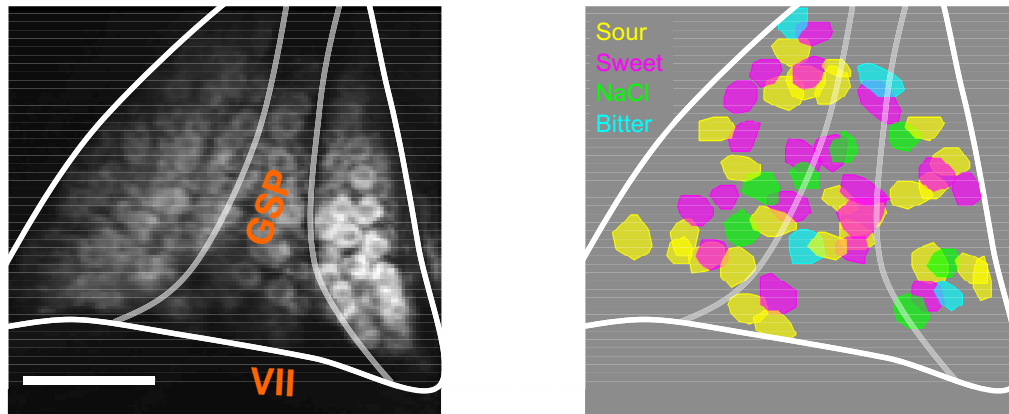
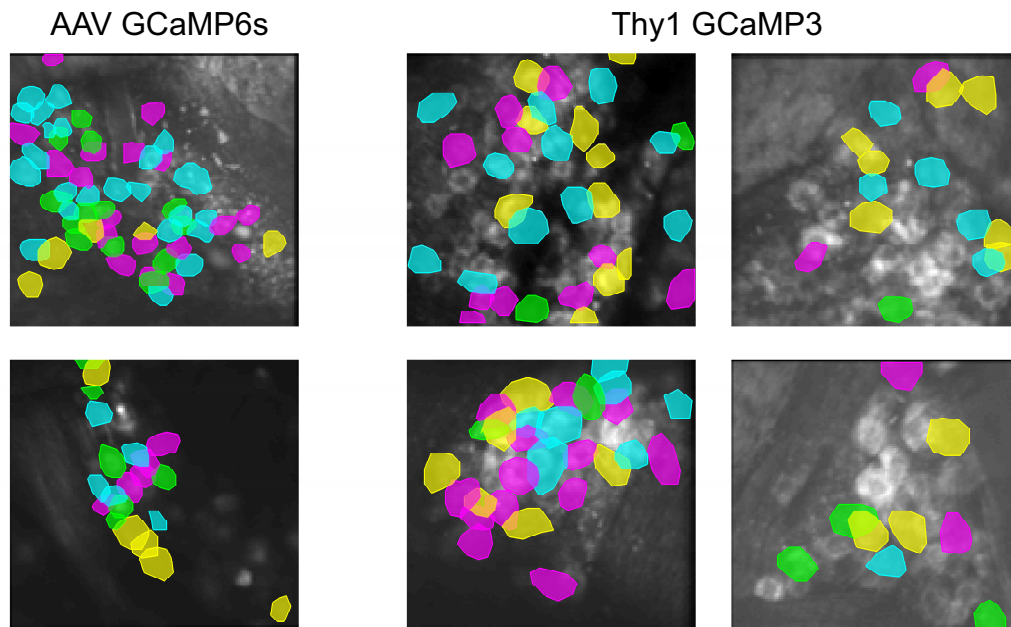
**Extended Data Figure 3 | Quantification of taste ganglion responses.**  
**a–d**, Rank-ordered plot of calcium transient amplitudes for various singly tuned ganglion neurons (see text and Fig. 3). For each cell, the mean response amplitudes for preferred stimulus (red) and the mean amplitude of its next-strongest tastant response (grey) are shown; minor dots indicate individual

trial amplitudes. **e**, Quantification of mean response amplitudes in singly tuned salt ganglion neurons before and after amiloride treatment (10  $\mu$ M,  $n = 23$  cells; paired  $t$ -test,  $P < 0.001$ ). **f**, Quantification of mean response amplitudes in singly tuned bitter cells before and after AITC treatment<sup>17</sup> (3 mM,  $n = 63$  cells; paired  $t$ -test,  $P < 0.001$ ).



**Extended Data Figure 4 | Bitter-sour ganglion cells receive taste information from bitter T2R-expressing cells.** Distribution of bitter, sour, and bitter-sour ganglion cells in a sample of control animals ( $n = 4$ ) and in animals expressing tetanus toxin in *PKD2L1*-expressing TRCs (*PKD2L1-TeNT*;  $n = 3$ ). As expected, no cells responsive to citric acid (50 mM) are detected in *PKD2L1-TeNT* mice<sup>17</sup>. However, bitter-sour cells are unaffected

(see Fig. 4), suggesting that activation of T2R-expressing TRCs mediates these acidic responses. As predicted, subsequent application of the bitter TRC inactivator AITC<sup>17</sup> abolishes bitter responses of the bitter ganglion neurons, as well as the bitter and sour responses in the bitter-sour cells. Note that the solid bars showing less than 1 cell are used to illustrate the lack of responding cells.

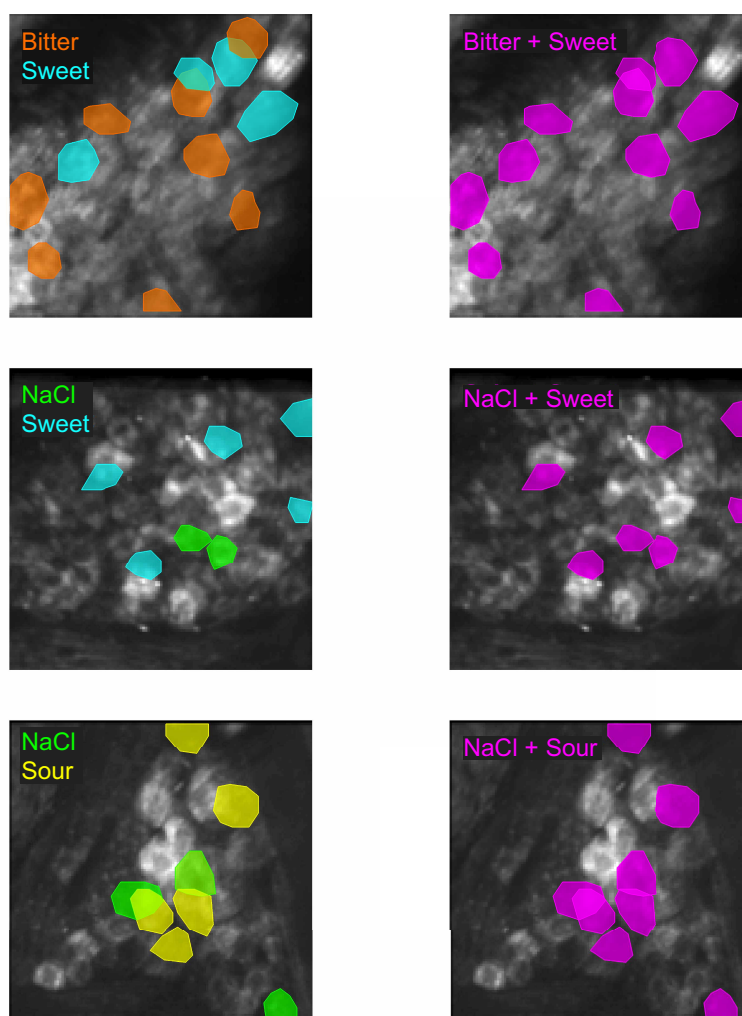
**a****b**

**Extended Data Figure 5 | Representation of taste quality does not cluster within the geniculate ganglion.** **a**, Two-photon endoscopic image (left) of a geniculate ganglion expressing GCaMP3. Highlighted are the locations of the facial (VII) and greater superficial petrosal (GSP) cranial nerves. The right panel shows approximately 50 neurons colour-coded according to their taste

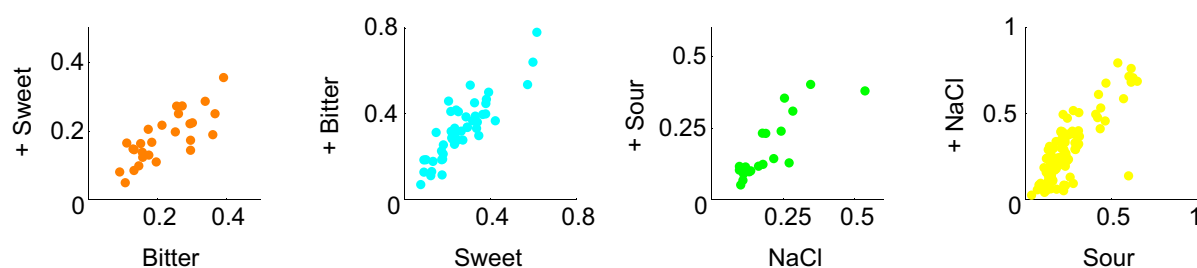
preference in this field. Sour, yellow; sweet, magenta; NaCl, green; bitter, cyan. Scale bar, 200  $\mu$ m. **b**, Representative fields of the geniculate ganglion from 6 different mice. The random distribution of neurons representing the various taste qualities is independent of sensor, or method of sensor delivery/expression (AAV-GCaMP6 or *Thy1*-GCaMP3); colour scheme same as for **a**.



a



b



**Extended Data Figure 6 | Representation of taste mixtures.** **a**, Imaging fields of three representative geniculate ganglia illustrating the ensembles of neurons recruited by two different single taste stimuli presented separately (left panels) versus the ensemble of neurons activated by a mixture of the two compounds presented together (right panels). See text for details; as expected there are no mixture-specific responders, and very few cells responded to each tastant in the mix: only 3 out of 113 cells examined with bitter + sweet responded to both tastants, 5 out of 301 cells examined with sour + salty responded to both, and 0 of 39 examined with salty + sweet responded to both tastants. We note that sour stimuli are known to suppress sweet responses<sup>30</sup>, but such suppression is sweet-cell autonomous and not due to interactions between

sweet and sour TRCs (data not shown). **b**, To quantitatively examine the impact of taste mixes on the responses of individual ganglion neurons, we analysed their response amplitudes in the presence of the single tastant versus the binary mix. Shown are plots of response amplitudes of a representative set of bitter, sweet, salty and sour geniculate neurons stimulated with their selective tastant ( $x$  axis) versus their response amplitude when in the presence of an additional tastant (as indicated in the  $y$  axis; shown are average  $\Delta F/F$  over 4 trials). 95% confidence intervals were determined using a ratio  $t$ -test: bitter + sweet/bitter, 0.73–0.91; sweet + bitter/sweet, 1.15–1.34; NaCl + sour/NaCl 0.74–1.00; sour + NaCl/sour, 0.95–1.16.

**Extended Data Table 1** | Shown are the numbers of responding neurons analysed in each of the 14 AAV-GCaMP6s and 15 Thy1-GCaMP3 mice

AAV-GCaMP6s

Animal	Responding Cells	Animal	Responding Cells	Animal	Responding Cells
1	70	2	57	3	51
4	28	5	25	6	25
7	24	8	21	9	19
10	17	11	12	12	10
13	10	14	9		

Thy1-GCaMP3

Animal	Responding Cells	Animal	Responding Cells	Animal	Responding Cells
1	91	2	66	3	31
4	30	5	26	6	26
7	24	8	24	9	22
10	21	11	21	12	18
13	15	14	8	15	8

**Extended Data Table 2 | The distribution of 971 taste ganglion neurons according to their responses to each of the five basic taste qualities**

Response profile	Number of Cells
NaCl	204
Bitter	168
Sour	152
Sweet	105
Umami	32
Bitter - sour	125
Sweet - umami	32
Sweet - NaCl	32
Sweet - sour	17
Sweet - bitter	14
Bitter - NaCl	12
NaCl - sour	11
NaCl - umami	8
Sweet - bitter - sour	8
Bitter - NaCl - sour	7
Sweet - NaCl - umami	6
Bitter - sour - umami	6
Sweet - sour - umami	5
Sweet - bitter - sour - umami	5
Bitter - umami	4
Sweet - bitter - umami	3
Sweet - NaCl - sour	3
Bitter - NaCl - sour - umami	3
NaCl - sour - umami	2
Sweet - bitter - NaCl	2
Sweet - bitter - NaCl - sour	2
Sweet - NaCl - sour - umami	2
Bitter - NaCl - umami	1

All the data showing responses in at least 1% of the total population (above horizontal line) are included in Fig. 4 (those below 1% were found on average in less than 1 in 4 animals). Note that bitter-sour-tuned neurons reflect the activation of T2R-expressing TRCs (see text and Methods for details).

# Control of plant stem cell function by conserved interacting transcriptional regulators

Yun Zhou<sup>1</sup>, Xing Liu<sup>1</sup>, Eric M. Engstrom<sup>2†</sup>, Zachary L. Nimchuk<sup>1,3†</sup>, Jose L. Pruneda-Paz<sup>4</sup>, Paul T. Tarr<sup>1</sup>, An Yan<sup>1</sup>, Steve A. Kay<sup>5</sup> & Elliot M. Meyerowitz<sup>1,3</sup>

Plant stem cells in the shoot apical meristem (SAM) and root apical meristem are necessary for postembryonic development of above-ground tissues and roots, respectively, while secondary vascular stem cells sustain vascular development<sup>1–4</sup>. WUSCHEL (WUS), a homeo-domain transcription factor expressed in the rib meristem of the *Arabidopsis* SAM, is a key regulatory factor controlling SAM stem cell populations<sup>5,6</sup>, and is thought to establish the shoot stem cell niche through a feedback circuit involving the CLAVATA3 (CLV3) peptide signalling pathway<sup>7</sup>. WUSCHEL-RELATED HOMEBOX 5 (WOX5), which is specifically expressed in the root quiescent centre, defines quiescent centre identity and functions interchangeably with WUS in the control of shoot and root stem cell niches<sup>8</sup>. WOX4, expressed in *Arabidopsis* procambial cells, defines the vascular stem cell niche<sup>9–11</sup>. WUS/WOX family proteins are evolutionarily and functionally conserved throughout the plant kingdom<sup>12</sup> and emerge as key actors in the specification and maintenance of stem cells within all meristems<sup>13</sup>. However, the nature of the genetic regime in stem cell niches that centre on WOX gene function has been elusive, and molecular links underlying conserved WUS/WOX function in stem cell niches remain unknown. Here we demonstrate that the *Arabidopsis* HAIRY MERISTEM (HAM) family of transcription regulators act as conserved interacting cofactors with WUS/WOX proteins. HAM and WUS share common targets *in vivo* and their physical interaction is important in driving downstream transcriptional programs and in promoting shoot stem cell proliferation. Differences in the overlapping expression patterns of WOX and HAM family members underlie the formation of diverse stem cell niche locations, and the HAM family is essential for all of these stem cell niches. These findings establish a new framework for the control of stem cell production during plant development.

To identify the molecular mechanism underlying WUS functions in stem cells, we screened for WUS-interacting transcription cofactors using yeast-two-hybrid assays with a transcription factor library<sup>14</sup>, and found that HAIRY MERISTEM 1 (HAM1) strongly and specifically interacts with WUS (Fig. 1a). HAM genes, encoding GRAS domain transcription regulators, contribute to shoot stem cell function in *Petunia* and *Arabidopsis*<sup>15–17</sup>. Four HAM genes (HAM1–HAM4) have been identified in *Arabidopsis*<sup>16</sup>, and further yeast assays revealed that WUS also interacted with three other HAM family members (Extended Data Fig. 1a). WUS–HAM associations were confirmed by bimolecular fluorescence complementation (BiFC) assays in tobacco (*Nicotiana benthamiana*), in which WUS and HAM were fused to the amino- and carboxy-terminal halves of green fluorescent protein (GFP), respectively (GFPn and GFPc). Strong GFP fluorescence in nuclei was observed when GFPn–WUS was co-transformed with GFPc–HAM (Fig. 1b, c and Extended Data Fig. 1b–e). WOX4 and WOX5 also interacted with HAM proteins in BiFC assays (Fig. 1d and Extended Data Fig. 1f–q). These WOX–HAM interactions were further confirmed through *in vitro* pull-down assays, where

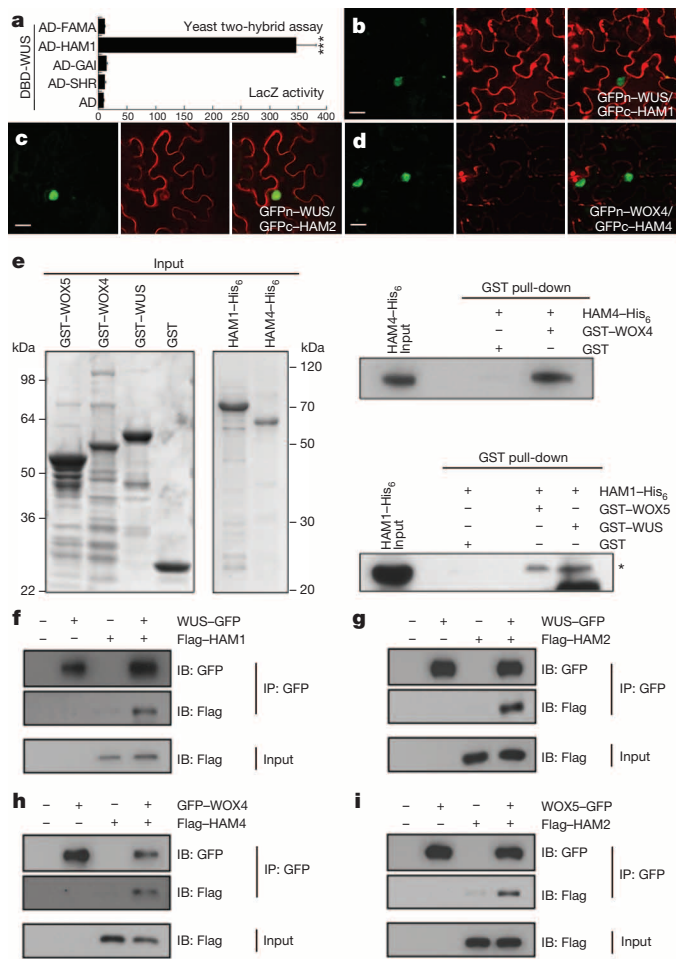
glutathione S-transferase (GST)–WOX4 but not GST bound HAM4–His<sub>6</sub>, and GST–WUS but not GST bound HAM1–His<sub>6</sub> (Fig. 1e). Interactions *in planta* were then tested using co-immunoprecipitation assays in tobacco, in which WUS–GFP bound Flag–HAM1 (Fig. 1f) and Flag–HAM2 (Fig. 1g), GFP–WOX4 bound Flag–HAM4 (Fig. 1h), and WOX5–GFP bound Flag–HAM2 (Fig. 1i). In short, with multiple approaches, our work revealed physical interactions between HAM and WUS/WOX family members.

We next constructed various deleted derivatives of HAM1 and WUS for yeast two-hybrid assays to identify essential regions for their interactions. Deleting amino acids from 117 to 230 (D117–230) in HAM1 abolished the interaction (Extended Data Fig. 2a). This amino-terminal fragment is important for HAM1 function in stem cell maintenance, as HAM1(D117–230) did not complement the *ham1;2;4* early termination phenotype, whereas full-length HAM1 driven by the same HAM1 promoter did (Extended Data Fig. 2b–g), and it is conserved in HAM proteins from *Arabidopsis* and across different plant species (Extended Data Fig. 2h–j). Deletion analyses of WUS identified a carboxy-terminal region required for interaction with HAM1 (Extended Data Fig. 3a), which is also required for WUS function (Extended Data Fig. 3b–d) and is conserved in different plant species (Extended Data Fig. 3e).

To dissect the roles of the HAM–WUS interaction in controlling shoot stem cell niches, genetic interactions were analysed between *ham1;2;3* (lacking the function of three of four HAM genes) and the weak *wus* allele *wus-7* (missense mutant), which forms a functional shoot apex<sup>18</sup> similar to wild type in terms of vegetative and inflorescence meristems (Fig. 2a, b, e). Different from *wus-7* single mutants (Fig. 2b) or *ham1;2;3* triple mutants (Fig. 2c), *wus-7;ham1;2;3* quadruple mutants display early termination of vegetative meristem development (Fig. 2d), thus resembling *wus* complete loss of function (null) mutants<sup>5</sup>. This effect also occurred in *wus-7;wus-7;ham1/ham1;ham2/ham2;ham3/+* plants, in which 41 out of 45 plants showed strong termination of inflorescence and floral meristems, with only leaves (Fig. 2h) or barren pedicels (flowers without carpels) (Fig. 2g) left at the top of the main shoot, a phenotype typical of *wus-1* null mutants<sup>5</sup>, but never observed in *wus-7* (Fig. 2e) or *ham1/ham1;ham2/ham2;ham3/+* (Fig. 2f) plants. Secondary inflorescence meristems initiated from axillary meristems in *wus-7;wus-7;ham1/ham1;ham2/ham2;ham3/+* plants also terminated prematurely (Extended Data Fig. 4a, b). Additionally, three out of four *wus-7;wus-7;ham1/ham1;ham2/ham2;ham4/+* plants displayed inflorescence meristem termination and lacked carpels (Extended Data Fig. 4c). A dose-dependent enhancement of stem cell termination was evident in *wus-7;ham1/+;ham2/+;ham3/+* and *wus-7;ham1/+;ham2/ham2;ham3/ham3* backgrounds (Extended Data Fig. 4d–f), demonstrating a functional interdependence between WUS and HAM family members *in vivo*. Downregulation of HAM1, HAM2 and HAM3 in a *ham4* shoot meristem, through activation of the microRNA MIR171—reported to target the HAM1, HAM2 and HAM3 genes<sup>19</sup>—led to terminated vegetative development (Extended

<sup>1</sup>Division of Biology, California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. <sup>2</sup>Biology Department, College of William and Mary, Williamsburg, Virginia 23187-8795, USA. <sup>3</sup>Howard Hughes Medical Institute, California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. <sup>4</sup>Section of Cell and Developmental Biology, Division of Biological Sciences, University of California San Diego, La Jolla, California 92093, USA. <sup>5</sup>University of Southern California, Molecular and Computational Biology, Department of Biological Sciences, Dana and David Dornsife College of Letters, Arts and Sciences, Los Angeles, California 90089, USA. <sup>†</sup>Present addresses: Department of Biology, Rollins College, Winter Park, Florida 32789, USA (E.M.E.); Department of Biological Sciences, Latham Hall RM 408, Virginia Tech, 220 Ag Quad Lane, Blacksburg, Virginia 24061, USA (Z.L.N.).

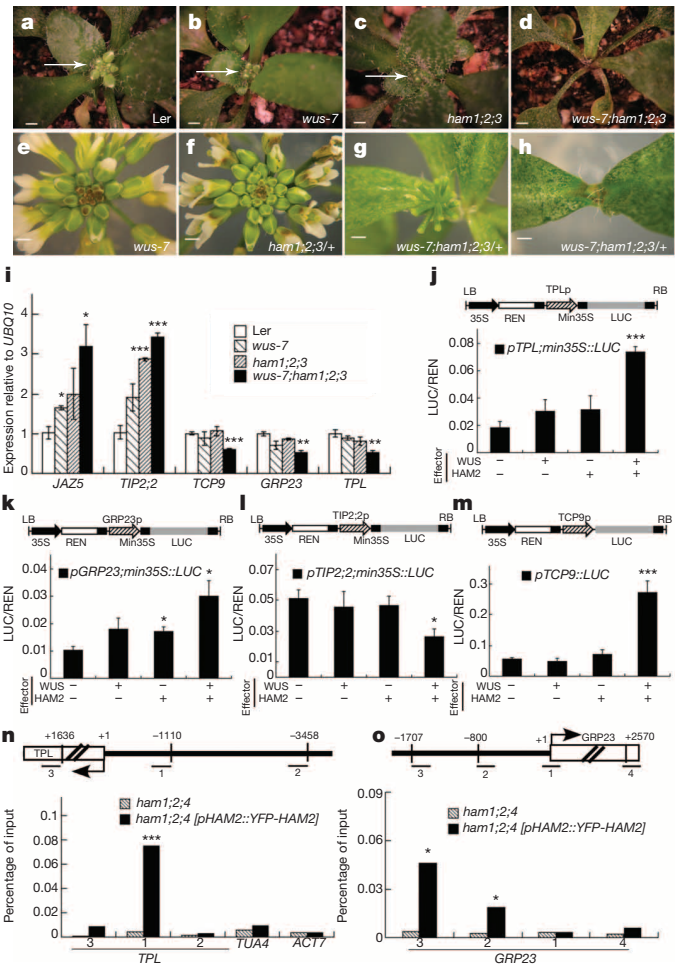




**Figure 1 | WUS/VOX and HAM family proteins physically interact.** **a**, LacZ activity in yeast two-hybrid assays. AD, activation domain; DBD, DNA-binding domain. Error bars show mean  $\pm$  standard error of the mean (s.e.m.) ( $n = 3$  biological replicates). \*\*\* $P < 0.001$  (two-tailed  $t$ -test). **b–d**, BiFC in tobacco. Panels (left to right): GFP; propidium iodide (PI) staining; merged channels. Scale bars, 20  $\mu$ m. **e**, SDS–polyacrylamide gel electrophoresis (SDS–PAGE) of input recombinant proteins stained by Coomassie blue (left), and pull-down of His<sub>6</sub>-tagged HAM proteins through GST-tagged WUS/VOX proteins detected by immunoblotting with anti-His antibody (right). Asterisk indicates HAM1–His<sub>6</sub> band and numbers indicate the apparent molecular weight of the protein bands in the protein standard. **f–j**, Co-immunoprecipitation of WUS–GFP and Flag–HAM1 (**f**), WUS–GFP and Flag–HAM2 (**g**), GFP–WOX4 and Flag–HAM4 (**h**), WOX5–GFP and Flag–HAM2 (**i**) (see Methods). IB, immunoblot; IP, immunoprecipitation.

Data Fig. 4g, h) similar to the *wus-1* phenotype, suggesting that WUS alone is not sufficient to maintain SAMs in the absence of HAM activity. Finally, the *wus-1;ham1;2;3* quadruple homozygote resembles a *wus-1* single mutant in several aspects including the vegetative meristem (Extended Data Fig. 4i–l), suggesting that WUS and HAM genes could act together at the SAM. All these genetic data are consistent with the hypothesis that WUS and HAM function as partners in shoot meristem maintenance.

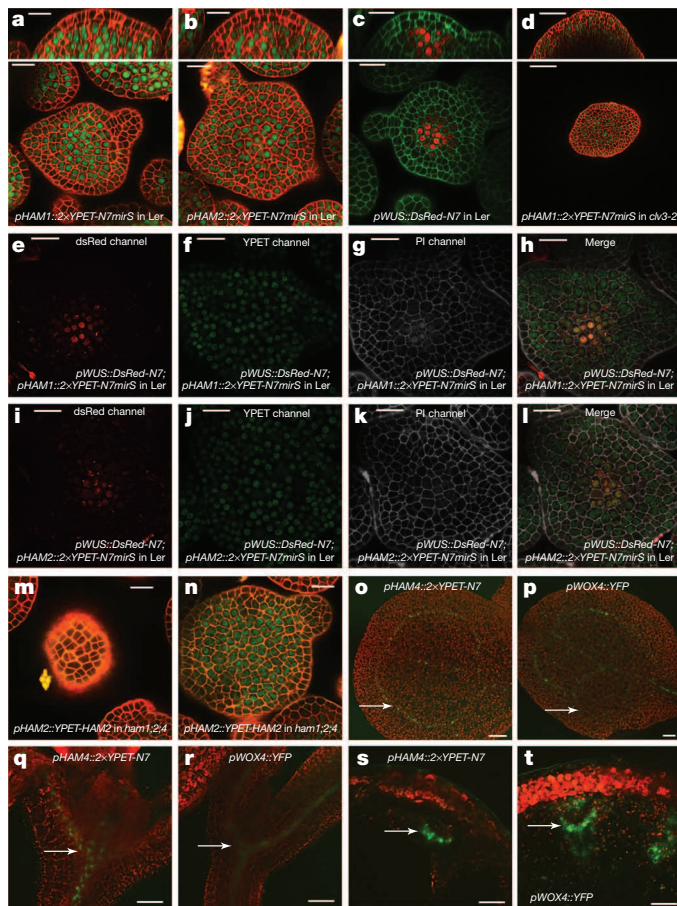
In addition to genetic interactions, the molecular function of the WUS–HAM interaction was further investigated. First, quantitative PCR with reverse transcription (RT–PCR) results (Fig. 3i) demonstrated that HAM proteins regulate expression of a set of genes including *JAZ5*, *TIP2;2*, *TCP9*, *GRP23* and *TPL*, which were reported to be directly regulated by WUS<sup>20</sup>. These WUS downstream targets were misregulated in *wus-7* or *ham1;2;3* triple mutants in similar manners, and *wus-7* and *ham1;2;3* synergistically regulated their expression (Fig. 2i), consistent with functional physical (Fig. 1) and genetic (Fig. 2a–h) interactions between WUS



**Figure 2 | WUS and HAM family genes cooperatively control the shoot stem cell niche and co-regulate a common gene set.** **a–d**, Shoot apices (a–d) (arrows) and inflorescence structures (e–h) of plants of indicated genotypes (Ler, wild type). Scale bars, 2 mm. **i**, RT–PCR quantification of WUS and HAM target gene expression in indicated genotypes. Error bars show mean  $\pm$  s.e.m. ( $n = 3$  biological replicates). **j–m**, Ratio of firefly luciferase (LUC) to *Renilla* luciferase (REN) activity in tobacco cells co-transformed with different reporter constructs (structure above each graph) and indicated effectors (see Methods). Min35S, 60-base-pair 35S minimum element; LB, transfer DNA (T-DNA) left border; RB, T-DNA right border. Error bars show mean  $\pm$  s.e.m. ( $n = 3$  biological replicates). **n**, **o**, ChIP of HAM2 protein with *TPL* or *GRP23* chromatin regions, with amplicon locations (bars with numbers) shown above each graph. The ChIP experiments were repeated three times using independent biological replicates with similar results, and one representative data set is shown. **i–o**, \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  (two-tailed  $t$ -test).

and HAM. Second, dual luciferase assays were conducted *in planta* to confirm the direct effects of WUS–HAM on target gene expression. Compared with empty-vector controls, the target genes examined were moderately (Fig. 2j–k) or barely (Fig. 2l, m) regulated by WUS or HAM alone, but were markedly affected when WUS and HAM were combined (Fig. 2j–l), indicating a role for the WUS–HAM interaction in regulating their transcription activities. Last, chromatin immunoprecipitation (ChIP) experiments demonstrated an *in vivo* association of yellow fluorescent protein (YFP)–HAM2 proteins with *TPL* (Fig. 2n) and *GRP23* promoters (Fig. 2o), genomic regions similar to those reported to associate with WUS protein *in vivo*<sup>20</sup>, supporting the notion that HAM family members are functional WUS cofactors in controlling the shoot stem cell niche through regulation of common target genes.

Consistently with physical and genetic interactions between HAM and WOX members, visualization of HAM and WUS/VOX fluorescent



**Figure 3 | HAM and WUS/WOX expression domains overlap.**

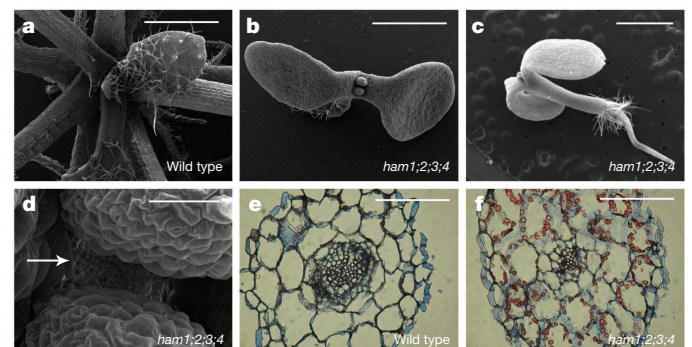
**a–d**, Expression of *pHAM1::2xYPET-N7MIRNASSENSITIVE* marker (*pHAM1::2xYPET-N7mirS*) (green) (**a**), *pHAM2::2xYPET-N7mirS* (green) (**b**), *pWUS::DsRed-N7* (red) (**c**) in Ler inflorescence meristem, and *pHAM1::2xYPET-N7mirS* (green) marker in a *clv3-2* inflorescence meristem (**d**). Orthogonal (top) and transverse section (bottom) views of the same plant are shown. **e–l**, Overlapping expression patterns of *pWUS::DsRed-N7* with *pHAM1::2xYPET-N7mirS* or *pHAM2::2xYPET-N7mirS* in the same shoot meristems (see Methods). Panels (from left to right): dsRed (red); YPET, an improved version of YFP (green); PI (grey); merged channels. **m, n**, Expression of *pHAM2::YPET-HAM2* translational marker (green) in L1 (**m**) and L3 (**n**) of the same *ham1;2;4* SAM. **o–t**, Overlapping expression patterns of *pHAM4::2xYPET-N7* and *pWOX4::YFP* (green, arrows) in the provascular and procambium cells in cotyledons (**o, p**), seedlings (**q, r**), and stem transverse sections (**s, t**). PI counterstain: red (**a, b, d, m, n**); green (**c**); grey (**g, h, k, l**). Chlorophyll autofluorescence: red (**o–t**). Scale bars: 50  $\mu$ m (**d, s, t**); 200  $\mu$ m (**o**); 100  $\mu$ m (**p–r**); 20  $\mu$ m (**a–c, e–n**).

transcriptional reporters revealed that WOX and HAM family expression overlapped in *planta*. In vegetative (Extended Data Fig. 5c–h) and inflorescence meristems (Fig. 3a–c), *HAM1* and *HAM2* expression overlapped with that of *WUS* in the rib meristem. *HAM1* is expressed in the rib meristem and peripheral zone but not in the L1 or L2 layers of the central zone (Fig. 3a and Supplementary Video 1), while *HAM2* expression peaks within the centre of the rib meristem (Fig. 3b). Similarly to *WUS* (Extended Data Fig. 5a, b), *HAM1* is negatively controlled by *CLV* signalling, as *HAM1* is expressed throughout *clv3-2* meristems (Fig. 3d). We imaged the *WUS* and *HAM1* or *HAM2* reporters in the same SAMs (Fig. 3e–l). Although expressed broadly, signals from *HAM1* (Fig. 3f) or *HAM2* (Fig. 3j) overlap with *WUS* signals (Fig. 3e, i) in the same rib zone cells (Fig. 3h, l and Extended Data Fig. 5i–p). As the *WUS* protein has been reported to move in the SAM from its site of transcription in the rib domain<sup>21</sup>, the *WUS* and *HAM1/HAM2* interaction domain in SAMs could be broader than their transcriptional domain overlap. We also examined the *HAM2* translational reporter *pHAM2::YPET-HAM2*

in the *ham1;2;4* SAM (Fig. 3m, n and Extended Data Fig. 6), which completely complements the *ham1;2;4* triple mutant (Extended Data Fig. 6a–c), and it showed a pattern similar to the *HAM2* transcriptional reporter: a strong signal in the centre starting from L3 and low or no signal in the L1 layer (Fig. 3m, n and Extended Data Fig. 6d, e). Taken together, the co-localization of *WUS* and *HAM1/HAM2* in SAMs is consistent with functional *WUS*–*HAM1/HAM2* interactions (Figs 1 and 2).

*HAM4* and *WOX4* are co-expressed in the provascular or procambial cell types of various tissues (Fig. 3o–t and Extended Data Fig. 7). In stem transverse sections, *HAM4* is expressed specifically in the procambium, overlapping with *WOX4* expression, as well as with the *HAM3* and *HAM1* expression domains (Fig. 3s, t and Extended Data Fig. 7j–l). The tightly co-regulated spatial and temporal *HAM4* and *WOX4* expression patterns are consistent with a *WOX4*–*HAM4* interaction module (Fig. 1h). Both *HAM2* transcriptional and translational reporters (Extended Data Fig. 8) are expressed in root meristem cells including the quiescent centre, overlapping with the quiescent-centre-specific *WOX5* expression domain<sup>8</sup>, consistent with previous reports from cell-type-specific transcriptome analyses<sup>22,23</sup> and indicating the possibility of *WOX5*–*HAM2* interactions in roots. Our finding that both *WUS* and *WOX5* interact with *HAM2* may be partially accounted for by the fact that *WUS* and *WOX5* are interchangeable in controlling SAMs and root apical meristems<sup>8</sup>. Taken together, distinct and overlapping expression patterns of *HAM* and *WOX* members indicate that specific *HAM*–*WOX* pairs function within different stem cell niches throughout the plant.

To address the importance of the entire *HAM* family in the control of stem cell niches, we generated a *ham1;2;3;4* quadruple homozygous mutant. Compared with wild type, *ham1;2;3;4* plants displayed growth arrest at the early seedling stage, containing short roots and terminated shoots with two small leaf-like structures 26 days after germination (DAG) (Fig. 4a–c and Extended Data Fig. 9a–d); the shoot apices exhibited valley-like shapes at 26 DAG, lacking functional meristems (Fig. 4d); the hypocotyl transverse sections showed clear vascular defects, and the vascular bundles had reduced numbers of xylem vessels, fibres (dark-blue-stained) and phloem cells (red-stained), consistent with a reduction in the stem cell activity necessary for generating these cell types (Fig. 4e, f). Moreover, mid-veins in *ham1;2;3;4* leaf-like tissues did not differentiate but instead accumulated a dark-staining cell mass, resembling ground tissue cells (Extended Data Fig. 9e, f). This is similar to, but much stronger than, the reported *WOX4* RNA interference phenotype<sup>10</sup>. Root meristematic activity is also severely compromised in *ham* multiple mutants. The quiescent centre and columella stem cells (CSCs) in *ham1;2;3;4* mutants displayed enlarged and irregular shapes (Extended Data Fig. 9g, h) and, with incomplete penetrance, the CSCs in *ham1;2;3* mutants differentiate (Extended Data Fig. 9i–l), resembling reported defects in *wox5* mutants<sup>8</sup>. However, the root phenotype of *ham1;2;3* or



**Figure 4 | HAM family members are essential for various plant stem cell activities.** Scanning electron microscopic imaging of wild-type (**a**) and *ham1;2;3;4* (**b–d**) seedlings (26 DAG). Arrow indicates a *ham1;2;3;4* plant lacking a functional SAM. **e, f**, Transverse sections of wild-type and *ham1;2;3;4* hypocotyls (7 DAG). Scale bars: 1 mm (**a–c, e, f**); 50  $\mu$ m (**d**).



*ham1;2;3;4* plants is much more severe than that of the *wox5* mutant, suggesting that HAM regulates root meristem development not only through direct interaction with WOX5 but also through WOX5-independent pathways. In summary, in diverse meristems, *ham1;2;3;4* mutants display defects that share similarities with mutants lacking WOX activities, supporting the idea that HAM proteins are cofactors for WUS/WOX-family-mediated stem cell niche maintenance. Given the evolutionary conservation of plant meristem cell niches and the WOX/HAM gene families<sup>12,16</sup>, and the fact that WOX–HAM interactions exist in flowering plants besides *Arabidopsis* (Extended Data Fig. 10), this work establishes a new basis for studying stem cell niches in *Arabidopsis*, and provides a paradigm for meristem cell control regimes likely to be universal in flowering plants.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 February; accepted 10 September 2014.

Published online 26 October 2014.

- Meyerowitz, E. M. Genetic control of cell division patterns in developing plants. *Cell* **88**, 299–308 (1997).
- Sablowski, R. The dynamic plant stem cell niches. *Curr. Opin. Plant Biol.* **10**, 639–644 (2007).
- Miyashima, S., Sebastian, J., Lee, J. Y. & Helariutta, Y. Stem cell function during plant vascular development. *EMBO J.* **32**, 178–193 (2012).
- Dinneny, J. R. & Benfey, P. N. Plant stem cell niches: standing the test of time. *Cell* **132**, 553–557 (2008).
- Laux, T., Mayer, K. F., Berger, J. & Jurgens, G. The *WUSCHEL* gene is required for shoot and floral meristem integrity in *Arabidopsis*. *Development* **122**, 87–96 (1996).
- Mayer, K. F. *et al.* Role of *WUSCHEL* in regulating stem cell fate in the *Arabidopsis* shoot meristem. *Cell* **95**, 805–815 (1998).
- Schoof, H. *et al.* The stem cell population of *Arabidopsis* shoot meristems is maintained by a regulatory loop between the *CLAVATA* and *WUSCHEL* genes. *Cell* **100**, 635–644 (2000).
- Sarkar, A. K. *et al.* Conserved factors regulate signalling in *Arabidopsis thaliana* shoot and root stem cell organizers. *Nature* **446**, 811–814 (2007).
- Hirakawa, Y., Kondo, Y. & Fukuda, H. TDIF peptide signaling regulates vascular stem cell proliferation via the *WOX4* homeobox gene in *Arabidopsis*. *Plant Cell* **22**, 2618–2629 (2010).
- Ji, J. *et al.* WOX4 promotes procambial development. *Plant Physiol.* **152**, 1346–1356 (2010).
- Suer, S., Agusti, J., Sanchez, P., Schwarz, M. & Greb, T. WOX4 imparts auxin responsiveness to cambium cells in *Arabidopsis*. *Plant Cell* **23**, 3247–3259 (2011).
- Nardmann, J., Reisewitz, P. & Werr, W. Discrete shoot and root stem cell-promoting WUS/WOX5 functions are an evolutionary innovation of angiosperms. *Mol. Biol. Evol.* **26**, 1745–1755 (2009).
- van der Graaff, E., Laux, T. & Rensing, S. A. The WUS homeobox-containing (WOX) protein family. *Genome Biol.* **10**, 248 (2009).
- Pruneda-Paz, J. L. *et al.* A genome-scale resource for the functional characterization of *Arabidopsis* transcription factors. *Cell Rep.* **8**, 622–632 (2014).
- Stuurman, J., Jaggi, F. & Kuhlemeier, C. Shoot meristem maintenance is controlled by a GRAS-gene mediated signal from differentiating cells. *Genes Dev.* **16**, 2213–2218 (2002).
- Engstrom, E. M. *et al.* *Arabidopsis* homologs of the petunia hairy meristem gene are required for maintenance of shoot and root indeterminacy. *Plant Physiol.* **155**, 735–750 (2011).
- Schulze, S., Schafer, B. N., Parizotto, E. A., Voinnet, O. & Theres, K. *LOST MERISTEMS* genes regulate cell differentiation of central zone descendants in *Arabidopsis* shoot meristems. *Plant J.* **64**, 668–678 (2010).
- Graf, P. *et al.* *MGOUN1* encodes an *Arabidopsis* type IB DNA topoisomerase required in stem cell regulation and to maintain developmentally regulated gene silencing. *Plant Cell* **22**, 716–728 (2010).
- Llave, C., Xie, Z., Kasschau, K. D. & Carrington, J. C. Cleavage of *Scarecrow-like* mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* **297**, 2053–2056 (2002).
- Busch, W. *et al.* Transcriptional control of a plant stem cell niche. *Dev. Cell* **18**, 841–853 (2010).
- Yadav, R. K. *et al.* WUSCHEL protein movement mediates stem cell homeostasis in the *Arabidopsis* shoot apex. *Genes Dev.* **25**, 2025–2030 (2011).
- Nawy, T. *et al.* Transcriptional profile of the *Arabidopsis* root quiescent center. *Plant Cell* **17**, 1908–1925 (2005).
- Brady, S. M. *et al.* A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* **318**, 801–806 (2007).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The authors are grateful to R. Deshaies for his support with the protein purification and pull-down experiments, to D. Rees for sharing the 96-well format luminometer, to T. Laux, T. Greb and X. Deng for sharing published reagents, to K. Sugimoto and A. Roeder for help with the histology experiments and critical reading of the manuscript, to A. Sampathkumar for the suggestion of confocal imaging, and to A. Garda and L. Wang for technical support. Scanning electron microscopy was performed at the Applied Research Center of the College of William and Mary with technical assistance from B. Robertson. This work was funded by National Institutes of Health (NIH) grant R01 GM104244 and by the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation (through grant GBMF3406) to E.M.M., by a Caltech Gosney Postdoctoral Fellowship to Y.Z., by NIH grants GM094212, GM056006 and GM067837 to S.A.K., and was aided by a grant from The Jane Coffin Childs (JCC) Memorial Fund for Medical Research to X.L., a JCC fellow.

**Author Contributions** Y.Z. and E.M.M. conceived the experiments. Y.Z., X.L., E.M.E. and A.Y. performed experiments. J.L.P.-P. and S.A.K. provided the transcription factor library. Z.L.N. and P.T.T. contributed reagents. Y.Z., X.L. and A.Y. analysed data. Y.Z. and E.M.M. wrote the manuscript and X.L., Z.L.N. and A.Y. revised it. All authors read and approved the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.M.M. ([meyerow@caltech.edu](mailto:meyerow@caltech.edu)).

## METHODS

**Plant materials and growth conditions.** *Arabidopsis thaliana* plants were grown in a sunshine soil/vermiculite/perlite mixture under continuous light at 20 °C. The mutant lines *ham1;2;3* (triply homozygous for mutant alleles of *ham1-1*, *ham2-1* and *ham3-1*), *ham1;2;4* (triply homozygous for mutant alleles of *ham1-1*, *ham2-1* and *ham4-1*), *wus-7*, *wus-1*, *clv3-2* were previously described<sup>5,16,18,24</sup>. *wus-7;ham1;2;3*, *wus-7;ham1;2;4*, *wus-1;ham1;2;3*, and *ham1;2;3;4* mutants were generated through genetic crosses, and identified based on PCR genotyping in the F2 segregating population. Different mutant combinations in an *er* background were chosen for genetic and morphological analyses. All of the phenotypes were confirmed from multiple independent segregation lines to control for differences in ecotype background. PCR genotyping was performed as previously described<sup>16,18</sup>. Reporter lines for *pWUS::DsRed-N7* and *pWOX4::YFP* were previously reported<sup>11,25</sup>.

**Yeast two-hybrid assay.** Yeast transformation and  $\beta$ -galactosidase assays were performed following the manufacturer's instructions (Clontech). Full-length cDNAs for *WUS*, *HAM1*, *HAM2*, *HAM3* and *HAM4* were cloned into pENTR/D/TOPO or pCR8 (Invitrogen), and then *WUS* cDNA was Gateway cloned to pDEST32, and *HAM1*, *HAM2*, *HAM3* and *HAM4* cDNAs were Gateway cloned into pDEST22 using standard LR reactions (Invitrogen). All of the deletion derivatives for *WUS* or *HAM1* were generated through overlapping PCR with the primers listed later, cloned into pENTR/D-TOPO or pCR8, and cloned into pDEST32 or pDEST22 through LR recombination (Invitrogen). All clones were sequenced to confirm that they were in-frame and with designed deletions before being transformed into yeast. The bait and prey vectors were transformed into yeast strain MaV203, and three single transformed colonies per genotype were used as triplicate for the LacZ liquid assay in 96 Deepwell plates (Thermo) and optical density (OD) readings were recorded in a 96-well plate reader (Tecan). LacZ activity was calculated as (OD<sub>420</sub> nm × 1,000)/(OD<sub>600</sub> nm × cell volume in  $\mu$ l × assay time in minutes) following the yeast two-hybrid handbook (Clontech), including a standard error from three biological replicates.

**BiFC.** For BiFC experiments, full-length *Arabidopsis WUS*, *WOX4*, *WOX5*, *HAM1*, *HAM2*, *HAM3*, *HAM4*, *BARD1* and *FAMA* cDNA Gateway clones were recombined into vectors containing each half of GFP (N or C terminus) to generate the fusion proteins (GFPn-WUS, GFPn-WOX4, GFPn-WOX5, GFPn-BARD1, GFPn-FAMA, GFPc-HAM1, GFPc-HAM2, GFPc-HAM3, GFPc-HAM4, GFPc-BARD1, GFPc-FAMA) as previously described<sup>26</sup>. Plasmid pairs for testing the specific interactions (such as GFPn-WUS and GFPc-HAM1) were co-transformed together with the P19 silencing suppressor<sup>27</sup> into *N. benthamiana* leaves through *Agrobacterium* infiltration. The infiltrated tobacco leaves were stained with PI and imaged using a Zeiss LSM 510 Meta confocal microscope two days after infiltration. Green GFP signals in nuclei (which demonstrate the physical interaction) and red PI staining signals (which indicate tobacco cell structure) were captured at the same time from different detection channels. A 488 nm laser line was used to stimulate GFP and PI. A 505–530 bandpass filter was used to collect GFP signal and a 585–615 bandpass filter was used to collect PI signal. BARD1, a nuclear-localized protein, was included as a negative control. FAMA, a bHLH transcription factor that has been demonstrated to interact with bHLH transcription factors<sup>28</sup>, was used as an additional negative control. The positive signals for each pair were confirmed with four independent biological replicates, and representative images are shown in the figures. The same method was also used for tomato (*Solanum lycopersicum*) proteins, including GFPn-tomato WUS, GFPn-tomato WOX4 and GFPc-tomato HAM.

**Co-immunoprecipitation and western blot analysis.** *WUS* or *WOX5* cDNA in pCR8 was recombined to pMDC83 (ref. 29) to generate a *WUS*-GFP or *WOX5*-GFP fusion clone. Flag-HAM1, Flag-HAM2 and Flag-HAM4 were PCR amplified with primers 5'-CACCATGgactacaaggacgacgatgacaaggcggtggaagtCCCTTATCCTTTGAAAGGTTTCAAGG-3', 5'-CTAACATTTCCAAGCAGAGACA GTAACAAGTTC-3', and with primers 5'-CACCATGgactacaaggacgacgatgacaaggcggtggaagtCCCTTGGCCTTTGAGCAATTT-3', 5'-TTAACATTTCCAAGCTGAGACAGTA-3', and with primers 5'-CACCATGgactacaaggacgacgatgacaaggcggtggaagtAAAATCCCTGCATCATCTCCTC-3', 5'-CTAAAACCGCCAAGCTGATGTGGCAACAAG-3', respectively (lower-case letters indicate coding sequences for Flag and a linker). GFP DNA was amplified and sub-cloned in front of *WOX4* cDNA in-frame to generate a GFP-WOX4 fragment. Flag-HAM1, Flag-HAM2, Flag-HAM4 and GFP-WOX4 were then recombined into pMDC32 (ref. 29). For co-immunoprecipitation of *WUS*-GFP with Flag-HAM1, *WUS*-GFP with Flag-HAM2, GFP-WOX4 with Flag-HAM4, or *WOX5*-GFP with Flag-HAM2 in *N. benthamiana*, the constructs were introduced into *N. benthamiana* leaves through *Agrobacterium* infiltration. The leaves were harvested 2 days after infiltration and frozen in liquid nitrogen. For the immunoprecipitation of YFP-HAM2 in *Arabidopsis*, the shoot apices from the transgenic plants *pHAM2::YFP-HAM2* in *ham1;2;4* were harvested. The nuclei from *Arabidopsis* or tobacco were isolated, and then lysed with RIPA buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 3 mM dithiothreitol (DTT), 2 mM NaF and 1 mM NaVO<sub>3</sub>,

or 50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS for co-immunoprecipitation of GFP-WOX4 with Flag-HAM4) containing protease inhibitor cocktail (Roche) and 200  $\mu$ M PMSF by incubation on ice for 30 min followed by brief sonication. Clear lysates were mixed with diluting buffer containing PMSF and protease inhibitor cocktail (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 3 mM DTT, 2 mM NaF and 1 mM NaVO<sub>3</sub>, or 50 mM Tris-HCl, pH 8.0, 60 mM NaCl for co-immunoprecipitation of GFP-WOX4 with Flag-HAM4) (1:5, v:v), immunoprecipitated with GFP-Trap agarose beads (ChromoTek), and the beads were washed three times with the diluting buffer in spin columns (BioRad). The recovered proteins were eluted from the beads by boiling in 2× SDS sample buffer, separated by SDS-PAGE and transferred to nitrocellulose membrane (Millipore). Proteins were detected using anti-GFP antibody (Roche, catalogue #11814460001), anti-Flag antibody (Sigma, catalogue #F1804), and horseradish peroxidase (HRP)-conjugated anti-mouse antibody (Promega, catalogue #W4021). The co-immunoprecipitation experiments were repeated twice with similar results.

**Protein expression constructs and protein purification.** *WOX4* cDNA was amplified with primers 5'-CATAGAATTTCATGAAGGTTTCATGAGTTTTCGAA-3' and 5'-AGTTGCGGCCGCTCATCTCCCTCAGGATGGAGAGGA-3' (restriction enzyme sites are in bold), and cloned in-frame in pGEX-4T-1 with EcoRI and NotI sites. *WOX5* cDNA was amplified with primers 5'-ATTTCGCGGGTATGTCTTTCTCCGTGAAAGGTCG-3' and 5'-AGTTGCGGCCGCTTAAAGAAAGCTTAATCGAAGATCT-3' (restriction enzyme sites are in bold), and cloned in-frame in pGEX-4T-1 with XmaI and NotI sites. *WUS* cDNA was amplified with primers 5'-CATAGAATTTCATGGAGCCGCCACAGCATCAG-3' and 5'-AGT TGCGGCCGCTAGTTCAGAGCTAGCTCAAGA-3' (restriction enzyme sites are in bold), and cloned in-frame in pGEX-4T-1 with EcoRI and NotI sites. HAM1-His<sub>6</sub> tag was generated from PCR with primers 5'-CATAGAATTTCATGCCCTTATCCTTTGAAAGGTTTCAAGG-3' and 5'-AGTTGCGGCCGCTAGTGA TGATGATGATGATGACATTTCCAAGCAGAGACAGTAACAAGTTCTT-3' (restriction enzyme sites are in bold, and His<sub>6</sub> coding sequence is underlined), and cloned in-frame with thrombin cutting site in pGEX-4T-1 with EcoRI and NotI. HAM4-His<sub>6</sub> tag was generated from PCR with primers 5'-CATAGAATTTCATG AAAATCCCTGCATCATCTCCTC-3' and 5'-AGTTGCGGCCGCTAGTGA TGATGATGATGATGAAACCGCCAAGCTGATGTGGCAACAAG-3' (restriction enzyme sites are in bold, and His<sub>6</sub> coding sequence is underlined), and cloned in-frame with thrombin cutting site in pGEX-4T-1 with EcoRI and NotI. All proteins were expressed in Rosetta *Escherichia coli* (Novagen) by inducing with 0.4 mM isopropyl- $\beta$ -D-thiogalactoside (IPTG) at 16 °C for 2–4 h. GST-WOX4, GST-WOX5 and GST were purified on glutathione resin. HAM1 was purified on glutathione resin followed by digestion with thrombin and chromatography on S200 resins as described previously<sup>30,31</sup>. HAM4 was purified on glutathione resin followed by digestion with thrombin and removal of the GST associated with glutathione resin as described previously<sup>31</sup>.

**In vitro pull-down assay.** GST-WOX4, GST-WOX5, GST-WUS or GST were immobilized on glutathione resin and incubated with HAM1-His<sub>6</sub> or HAM4-His<sub>6</sub> for 30 min at 4 °C. The glutathione resin was then washed three times and processed for SDS-PAGE analysis and western blot analysis using antibody to His-tag (Qiagen, catalogue #34660). The pull-down experiments were repeated twice with similar results.

**Transactivation assay in tobacco.** A 60 base pair (bp) minimal 35S fragment (the –60 minimal promoter) was amplified and cloned with BamHI/NcoI sites into the pGREEN800II LUC<sup>32</sup> to generate a pGREEN800II-60LUC. *TPL* promoter was PCR amplified from Col-0 genomic DNA with primers 5'-AACAGGTACCGAAGCGTTCGTTTCATTAGTTTATC-3' and 5'-AATAAGGATCCGTTTCTCTCACTTCCTTAAAGACT-3' (restriction enzyme sites are in bold) and cloned with KpnI/BamHI sites into the pGREEN800II-60LUC. *TIP2* promoter was amplified with primers 5'-AACAGGTACCGAGTGAAGCAGATGGGAGAGAA-3' and 5'-AATACTGCAGTTGATCCGACAAAATAACTCTGTT-3' and cloned with KpnI/PstI sites into the pGREEN800II-60LUC. *GRP23* promoter was amplified with primers 5'-AACAGGTACCGAGGTGTGATTGTCAATAGACTACG-3' and 5'-AACAGATATCGGTGGAGGGAAATGATTTAGGGTT-3' and cloned with KpnI/EcoRV sites into the pGREEN800II-60LUC. *TCP9* promoter was amplified with primers 5'-AACAGGTACCGTATGCTGATGGTACGCAAAAGTT-3' and 5'-AATACTGCAGTAAAATATAGTGTGAGAGAAACG-3' and cloned with KpnI/PstI sites into the pGREEN800II LUC. The different reporter constructs (dual-luciferase reporter with different gene promoters) and indicated effectors (empty effector vector or *WUS* or *HAM2*, or *WUS* together with *HAM2*) were introduced into *N. benthamiana* leaves through *Agrobacterium* infiltration. The activities of LUC and REN were quantified 2 days after infiltration with a Dual Luciferase Assay kit (Promega), and luminescence was recorded using a 96-well dual injection luminometer (Tecan). The LUC activity was normalized to the REN activity (LUC/REN). The means and standard errors of LUC/REN were calculated from three independent biological replicates.



**Plasmid constructions for the transgenic plants.** It has been previously reported that *HAM1*, *HAM2* and *HAM3* are targeted and repressed by the *MIR171* family<sup>19</sup>. To generate new microRNA-sensitive fluorescence reporters for *HAM1*, *HAM2* and *HAM3*, an approach similar to that in a previous report<sup>33</sup> was used. Briefly, a 2×*YPET-N7mirS* fragment was generated through PCR amplification, which contains a 2× version of YPET with a N7 nuclear localization sequence (2×*YPET-N7*) followed by 26 bp of microRNA target sequence (GCAAGGGATATTGGCGCGGCTCAATC) from the *HAM* family. These 26 bp are recognized and targeted by the *MIR171* family<sup>19,34</sup>.

For the construction of the *pHAM1::2×YPET-N7mirS* reporter, a 4 kb *Ascl* fragment containing the *HAM1* promoter was amplified from Col-0 genomic DNA with primers 5'-TACAGGCGCGCCTTCCCTCACTTTTCTTACATTGAA-3' and 5'-TACAGGCGCGCCACGCCTCTCAACAACACAGAGTAA-3' (restriction enzyme sites are in bold), and cloned 5' of the 2×*YPET-N7mirS* fragment. The fused DNA fragment was introduced into the pMOA34 binary vector<sup>35</sup>. For the construction of *pHAM2::2×YPET-N7mirS*, the 3,122 bp *HAM2* promoter was amplified with 5'-TACAGTTTAAACAGCAGGACATATCTAAACCAGAGTT-3' and 5'-TACAGTTTAAACGACCAATCTTACAGAGTCAGAAAGA-3' (restriction enzyme sites are in bold) and cloned in front of 2×*YPET-N7mirS*; and the 1,149 bp *HAM2* 3' untranslated sequence was PCR amplified with 5'-TACAGGCGCGCCGACGAAAAAGGAGGATATTTTACCGGT-3' and 5'-TACAGGCGCGCCACTATGTTTCCATGTACTGTGGGATA-3' (restriction enzyme sites are in bold) and cloned 3' of the 2×*YPET-N7mirS* construct, then the fused DNA fragment was cloned into pMOA34. For the construction of *pHAM3::2×YPET-N7mirS*, the 3,816 bp *HAM3* promoter was amplified with 5'-TACAGTTTAAAC TTTATAAGACTTGCTATGGTCGTGAG-3' and 5'-TACAGTTTAAACTGCA GACGATAAAAAATAGTGTATT-3' (restriction enzyme sites are in bold) and cloned before 2×*YPET-N7mirS*; and the 1,755 bp *HAM3* 3' untranslated sequence was PCR amplified with 5'-TACAGGCGCGCCTTCCACCGGAGTTCAATT ATTTAA-3' and 5'-TACAGGCGCGCCTTAGTTGAAGGACAAATAACACCA AA-3' (restriction enzyme sites are in bold) and cloned 3' of the 2×*YPET-N7mirS* fragment, then the fused DNA fragment was introduced into pMOA34. The double reporter lines, including the *pWUS::DsRed-N7*; *pHAM1::2×YPET-N7mirS* line and the *pWUS::dsRed-N7*; *pHAM2::2×YPET-N7mirS* line, were generated through genetic crosses.

For the construction of the *pHAM4::2×YPET-N7* reporter, the 6,413 bp *HAM4* promoter was amplified with primers 5'-TACAGGCGCGCCAAATATAAAAT AGAATCAACAAAGTTGGTAAC-3' and 5'-CAAAAGGCGCGCGGTGTTGT GTGTTAAGAAGAAAGAGGTGGAGCCTTT-3' (restriction enzyme sites are in bold), and cloned 5' of a 2×*YPET-N7* fragment, then the fused DNA fragment was cloned into pMOA34.

For the complementation of *wus-1*, a full-length *WUS* or *WUS* derivative without base pairs encoding amino acids from 203 to 236 was cloned into the pMOA36 binary vector, together with 4.4 kb of the *WUS* upstream regulatory sequence and 1.5 kb of the *WUS* 3' untranslated sequence. The construct was introduced into *wus-1/+* plants using the floral dip method. For the complementation of *ham1;2;4*, *HAM1* or the *HAM1* derivative without 117–230 was cloned into the pMOA34 binary vector, with 3,949 bp of the *HAM1* upstream regulatory sequence and 1,387 bp of the *HAM1* 3' untranslated sequence. The construct was introduced into *ham1;2;4* plants using the floral dip method.

To generate a *MIR171* expression construct in shoot meristems, the *MIR171* DNA was amplified with 5'-CACCTGAGCGCACTATCGGACATCAAA-3' and 5'-TAAACGCGTGATATTGGCAC-3' and cloned into pMOA36 together with 4.4 kb of the *WUS* upstream regulatory sequence and 1.5 kb of the *WUS* 3' untranslated sequence. The construct was introduced into the *ham4* mutant through the floral dip method. Five independent transgenic plants (*pWUS::MIR171* in *ham4*) showing terminated vegetative meristems were identified.

**Confocal imaging of fluorescence reporters in living plants.** All of the fluorescent reporters were imaged by using a Zeiss LSM 510 Meta confocal microscope, except for the fluorescent reporters in inflorescence meristems and *HAM2* fluorescent reporters in the roots, which were imaged by using a Zeiss LSM 780 Meta confocal microscope. Zeiss LSM software was used for reconstructing the Z-stacks for a projection view. Laser and filter settings were used as described previously<sup>36–38</sup>. To image *HAM4* and *WOX4* reporters, the cotyledons, first leaf, hypocotyls and roots from 7-day-old seedlings and stems from 1-cm bolting plants were used. To image *dsRed*, *YPET* and *PI* simultaneously in SAMs, the multitracking mode in the ZEISS LSM 780 was used. *dsRed* was excited using a 561 nm laser line in conjunction with 571–589 nm collection; *YPET* was excited using a 514 nm laser line in conjunction with a 519–549 nm collection; and *PI* was excited using a 514 nm laser with 631–673 nm collection. There is no spectral bleed-through of *dsRed* into the *YPET* collection channel, nor of *YPET* into the *dsRed* collection channel under these settings, and for better display, all images from the *dsRed* channel were

equally enhanced with the same scale and all images from the *PI* channel were uniformly enhanced to the similar intensity using ImageJ software.

**Histology.** The wild-type and *ham1;2;3;4* seedlings were fixed in 4% paraformaldehyde, dehydrated and embedded in Paraplast X-tra (Fisher). The samples in wax were sectioned at 8 µm, de-waxed and dehydrated, and the slides were stained with Alcian blue together with Safranin O (red) as previously described<sup>39</sup>, to detect non-lignified cell walls and lignified cell walls, respectively.

**Real-time RT-PCR analysis.** Total RNA was isolated from 10-day-old plants with roots, hypocotyls and leaves dissected off, using the RNeasy Kit (Qiagen). SuperScript III reverse transcriptase (Invitrogen) was used to synthesize the first-strand cDNA with oligo(dT) primer and 1 µg of total RNA at 50 °C for 1 h. Quantitative PCR was then performed with the SensiMix SYBR Hi-ROX Kit (BioLine) on Roche Real-Time PCR machine following the manufacturer's instruction. The thermal cycling program was 95 °C for 10 min, followed by 45 cycles of 95 °C for 10 s, 56 °C for 30 s, 72 °C for 40 s, and a one-cycle dissociation stage at 95 °C for 15 s, 60 °C for 1 min, and 97 °C for 15 s. The primers used in quantitative RT-PCR were: *JAZ5*, 5'-GAAAGACAGAGCTGTGGCTAGG-3' and 5'-TTGGCCTTCTTCAATCTT CATAATA-3'; *TIP2*, 5'-ACCAATGGCGAGAGCGTACCG-3' and 5'-ATGA AACCGATAGCAATTGGAG-3'; *TCP9*, 5'-ACCTCCTTTACAAGTTGTTCAG-3' and 5'-TGAAGCTCTTGTTCCTGTATATCTC-3'; *GRP23*, 5'-AGACA GCTAGCCATCAGCAGTCAC-3' and 5'-AGTTCTCACTCCACTACCTTTT-3'; *TPL*, 5'-AGCTAGTCTCAGCAATTCAAA-3' and 5'-AGGCTGATCAG ATGACAGAGG-3'; and *UBQ10*, 5'-AACAAATGGAGGATGGTCTCGT-3' and 5'-T TCCAGGGAAGATGAGACG-3'. Fold change was calculated as 2<sup>ΔΔCt</sup> and standard error was calculated from three biological replicates, and each biological replicate was examined in triplicate.

**ChIP.** For the construction of *pHAM2::YFP-HAM2* (*pHAM2::YPET-HAM2*), the YFP variant YPET was amplified and cloned in front of *HAM2* cDNA in-frame to generate the YFP-*HAM2* fragment. Then the 3,122 bp *HAM2* promoter was amplified with 5'-TACAGTTTAAACAGCAGGACATATCTAAACCAGAGTT-3' and 5'-TACAGTTTAAACGACCAATCTTACAGAGTCAGAAAGAG-3' (restriction enzyme sites are in bold) and cloned in front of YFP-*HAM2*, and the 1,149 bp *HAM2* 3' untranslated sequence was amplified with 5'-TACAGGCGCGCCGAC GAAAAAGGAGGATATTTTACCGGT-3' and 5'-TACAGGCGCGCCACTAT GTTTCATGTACTGTGGGATA-3' and cloned 3' of YFP-*HAM2*. Then the whole fused DNA fragment (*pHAM2-YFP-HAM2-HAM2* 3' UTR) was cloned into the binary vector pMOA34. The construct was introduced into *ham1;2;4* plants using the floral dip method, and the complemented *ham1;2;4* [*pHAM2::YFP-HAM2*] line was selected for the western blot, GFP immunoprecipitation (shown in Extended Data Fig. 6f) and ChIP experiments.

A ChIP followed by a quantitative real-time PCR approach was used to investigate the *in vivo* association of *HAM2* with the *TPL* and *GRP23* promoters as described previously<sup>40</sup> with some modifications. In general, 2 g of *ham1;2;4* (negative control) or *ham1;2;4* [*pHAM2::YFP-HAM2*] plants were harvested and fixed with 1% formaldehyde under vacuum. Nuclei were isolated and lysed, and chromatin was sheared to an average size of 500 bp by sonication seven times for 20 s each with a Branson Sonifier. Samples were kept on ice during sonication and were cooled for 1 min between sonication pulses. The sonicated chromatin served as input. Immunoprecipitations were performed with GFP-Trap Agarose beads (Chromotek) at 4 °C following the manufacturer's procedure. The precipitated DNA was isolated and purified, and served as a template for PCR. Quantitative PCR was performed as described earlier. The relative enrichment for each immunoprecipitated amplicon (from *TPL* or *GRP23* promoter) from GFP-Trap is presented as ChIP/input ratio, and *TUA4* and *ACTIN7* (*ACT7*) amplicons are also included to serve as negative controls. The ChIP experiments were conducted three times using independent biological replicates with similar results, and one representative data set with two technical replicates is presented. The primer pairs used in ChIP-PCR are as follows: *TPL* amplicon 1, 5'-GCAATTGGCTCTTCAATGTC-3' and 5'-GGACGGAGAT CTAACGGCTA-3'; *TPL* amplicon 2, 5'-CCATATGACCGGGATATGAGA-3' and 5'-GGGATATGTCGCTTCCATT-3'; *TPL* amplicon 3, 5'-TTGAGTCAGG GCTCATCTCC-3' and 5'-CTTTCGCGGAGAACCACTTC-3'; *GRP23* amplicon 1, 5'-ACCATCGTCATTGGTTTCGT-3' and 5'-GGAGTGACTGAGAGACA TGG-3'; *GRP23* amplicon 2, 5'-CAACAAATCTCTGTTTTCACGTT-3' and 5'-C GAAAATGTTTGAAGTGCAT-3'; *GRP23* amplicon 3, 5'-CGCATCGCCTAA AAGTAA-3' and 5'-TTTGTGGCTAGGCATAGGG-3'; *GRP23* amplicon 4, 5'-AGACAGTTTATGCCATCAGCAGTCAC-3' and 5'-AGTTCTCACTCCACTCA CTACCTTTT-3'; *TUA4* amplicon, 5'-CTTTGGTCTTTAGCAGGTTTC-3' and 5'-CCCATCTGTATATAACGACAC-3'; *ACTIN7* amplicon, 5'-TGCTTGTTAT GTGATTCGATCC-3' and 5'-GATCGACAGAAGCGAGAAGAAT-3'.

**Staining.** mPS-PI staining and root imaging of the staining was performed as previously described<sup>41</sup>.

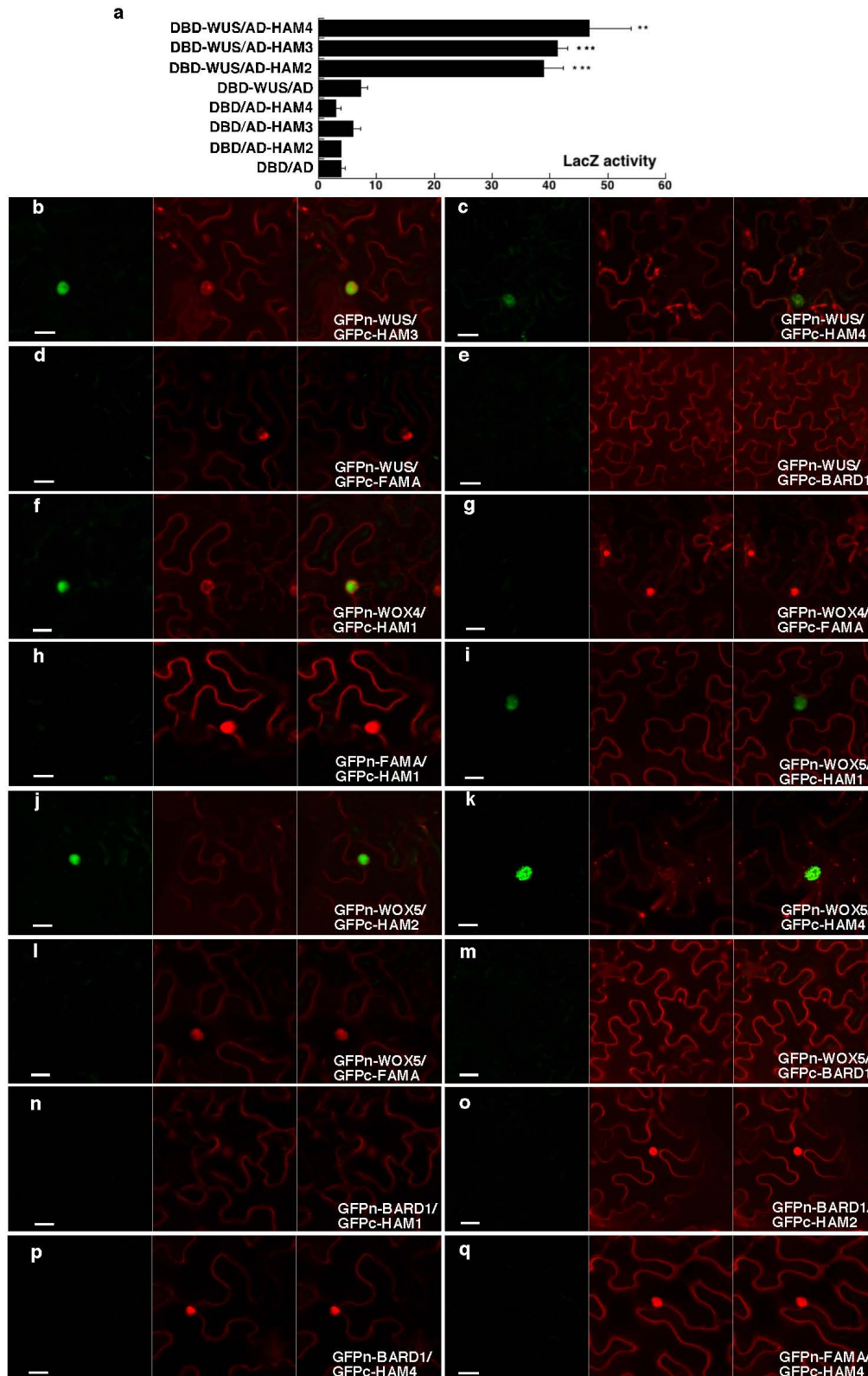
**Scanning electron microscopy.** For scanning electron microscopy, tissue was placed in 1.2% glutaraldehyde in 0.025 M phosphate buffer (sodium phosphate, pH 6.8),

vacuum was applied for 10 min, and tissue was fixed overnight at 4 °C. Tissue was then rinsed twice with 0.025 M phosphate buffer for 1 h, post-fixed with 0.5% osmium tetroxide in 0.025 M phosphate buffer for 24 h at room temperature, and moved through an increasing ethanol series (20% increments), each increment lasting a minimum of 1 h and ending with two exchanges of 100% ethanol. Ethanol was removed by critical point drying with a critical point drier (SAMDR1), and tissue was mounted to stubs with double-sided adhesive tape and sputter coated with gold-palladium alloy using a Hummer Sputtering System (Anatech). Samples were examined with a Hitachi 4700 scanning electron microscope.

**Primers used for cDNA clones and deletion constructions.** *HAM1c/5CACC*, 5'-CACCATGCCCTTATCCTTTGAAAGGTTTCAAGG-3'; *HAM1c/3*, 5'-ACA TTTCCAAGCAGAGACAGTAACAAG-3'; *HAM1c5/231*, 5'-CCGTTTTATCAC AACAACAG-3'; *HAM1c5/441*, 5'-GAAAATCTCAAAACATTCG-3'; *HAM1D71-116/5*, 5'-AGTCTCTCGCTTCTTATTCTGCTTCTTCTCTGGTCAAGAGC-3'; *HAM1D71-116/3*, 5'-GCTCTTGACAGGAGAGAAGCAGAATAAGAAGCG AGAGGACT-3'; *HAM1c71/5CACC*, 5'-CACCATGTCTACCACCACACGCT GTCTTCTCT-3'; *HAM1D117-230/5*, 5'-GATGATCTTGACGGTGTCTCT CTCCGTTTTATCACAACAACAGCAA-3'; *HAM1D117-230/3*, 5'-TTGCTGG TTGTGTGATAAAACGGAGAGAGAACCCTCAAGATCATC-3'; *HAM4c/5*, 5'-ATGAAAATCCCTGTCATCATCTCCTC-3'; *HAM4c/3*, 5'-AAACCGCCAA GCTGATGTGGCAACA-3'; *WUS5c/1*, 5'-ATGGAGCCGCCACAGCATCAG-3'; *WUS5c/30*, 5'-TACACGTGTGCGCAGACAGCAG-3'; *WUS5c/100*, 5'-AGATTCA ACGGAACAAACATGAC-3'; *WUS5c/171*, 5'-GCAAGCTCAGGTACTGAATG T-3'; *WUS5c/3stop*, 5'-CTAGTTCAGACGTAGCTCAAGA-3'; *WUS5c/236*, 5'-A CCTTCTAGACCAAACAGAGG-3'; *WUS5c/292*, 5'-GTTCTAGACGTAGCTCA AGAGAAGC-3'; *WUS1D164-183/5*, 5'-TAACAAGCCATATCCCAGCTTCAAT GGCTACATGAGTAGCCATG-3'; *WUS1D164-183/3*, 5'-CATGGCTACTCATG TAGCCATTGAAGCTGGGATATGGCTTGTTA-3'; *WUS1D101-163/5*, 5'-GGC TCGTAGCGTCAGAAGAAGAGAAATAACGGGAATTTAAATCATGCAA-3'; *WUS1D101-163/3*, 5'-TTGCATGATTTAAATTCCTGTTATTTCTCTTCTTCT GACGCTCACGAGCC-3'; *WUS1D132-163/5*, 5'-TATCATCTCTACTTCACC ATCATAAATAACGGGAATTTAAATCATGCAA-3'; *WUS1D132-163/3*, 5'-TT GCATGATTTAAATTCCTGTTATTTATGATGGTGAAAGTAGAGGATGATA-3'; *WUS1D184-236/5*, 5'-AATGTGGTGTGTTAATGCTTCTCATCAAGAAGAA GAAGAATGTGG-3'; *WUS1D184-236/3*, 5'-CCACATTCTTCTTCTTCTTGAT GAGAAGCATTAACAACACCACATT-3'; *WUS1D164-236/5*, 5'-TAACAAGC CATATCCCAGCTTCCATCAAGAAGAAGAAGATGTG-3'; *WUS1D164-236/3*, 5'-CACATTCTTCTTCTTCTTGATGGAAGCTGGGATATGGCTTGTTA-3'; *WUS1D184-202/5*, 5'-AATGTGGTGTGTTAATGCTTCTTACAACAACGTA GGTGGAGGAT-3'; *WUS1D184-202/3*, 5'-ATCCTCCACCTACGTTGTTGTA AGAAGCATTAACAACACCACATT-3'; *WUS2D203-236/5*, 5'-TGGAACAAGA CTGTTCTATGAATCATCAAGAAGAAGAAGATGTGG-3'; *WUS2D203-236/3*, 5'-CCACATTCTTCTTCTTCTTGATGATTCATAGAACAGTCTTGTTCCA-3'; *WUS2D218-236/5*, 5'-GGGCAAACATGGATCATCATTACCATCAAGAAGAA GAAGAATGTGG-3'; *WUS2D218-236/3*, 5'-CCACATTCTTCTTCTTCTTGAT GGTAATGATGATCCATGTTTGGCC-3'; *WUS2D203-217/5*, 5'-TGGAACAAG ACTGTTCTATGAATTCATCTGCACCTACAACTTCTT-3'; *WUS2D203-217/3*, 5'-AAGAAGTTGTAAGGTGCAGATGAATTCATAGAACAGTCTTGTTCCA-3';

*WOX4c/5CACC*, 5'-CACCATGAAGGTTTCATGAGTTTTCGAA-3'; *WOX4c/3stop*, 5'-TCATCTCCCTTCAGGATGGAGAGGA-3'; *WOX5c/5CACC*, 5'-CACCAT GTCTTTCTCCGTGAAAGGTC-3'; *WOX5c/3*, 5'-AAGAAAGCTTAATCGAA GATCT-3'; *TomatoHAM/5CACC*, 5'-CACCATGATTGTAATACCTCAAAGT AATAA-3'; *TomatoHAM/3stop*, 5'-TTAAAGAAAATCTCTTCTGGCTTCA GA-3'; *TomatoWUS/5CACC*, 5'-CACCATGGAACATCAACACAACATAGA AGA-3'; *TomatoWUS/3stop*, 5'-TTAGGGGAAAGAGTTGAGAGTAAGT-3'; *TomatoWOX4/5CACC*, 5'-CACCATGTACATGGGATCATCATCAGGAAG-3'; *TomatoWOX4/3stop*, 5'-TCATCTCATGCCTTCTGGATGCAATG-3'.

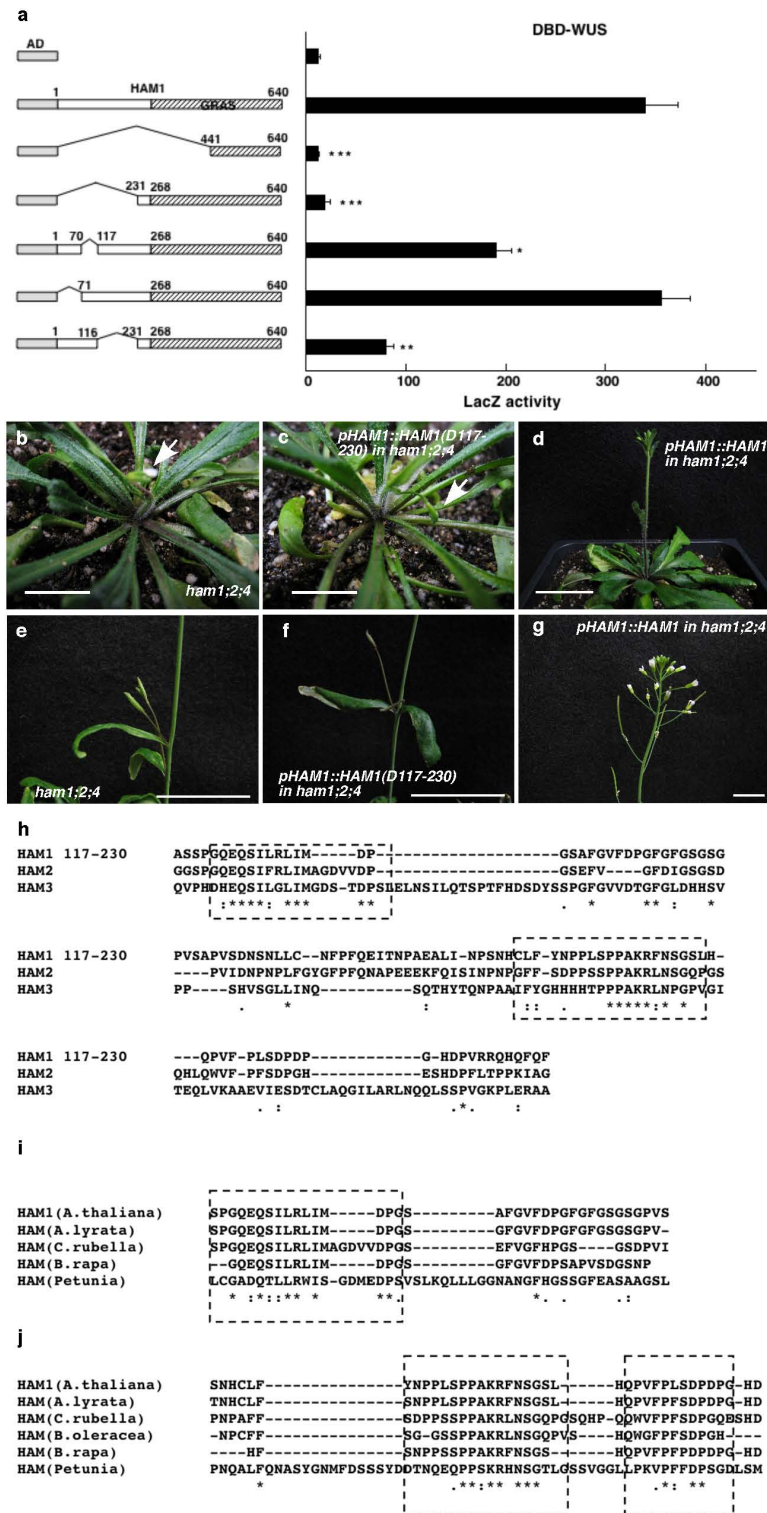
24. Fletcher, J. C., Brand, U., Running, M. P., Simon, R. & Meyerowitz, E. M. Signaling of cell fate decisions by CLAVATA3 in *Arabidopsis* shoot meristems. *Science* **283**, 1911–1914 (1999).
25. Gordon, S. P., Chickarmane, V. S., Ohno, C. & Meyerowitz, E. M. Multiple feedback loops through cytokinin signaling control stem cell number within the *Arabidopsis* shoot meristem. *Proc. Natl. Acad. Sci. USA* **106**, 16529–16534 (2009).
26. Walter, M. *et al.* Visualization of protein interactions in living plant cells using bimolecular fluorescence complementation. *Plant J.* **40**, 428–438 (2004).
27. Voinnet, O., Rivas, S., Mestre, P. & Baulcombe, D. An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. *Plant J.* **33**, 949–956 (2003).
28. Ohashi-Ito, K. & Bergmann, D. C. *Arabidopsis* FAMA controls the final proliferation/differentiation switch during stomatal development. *Plant Cell* **18**, 2493–2505 (2006).
29. Curtis, M. D. & Grossniklaus, U. A gateway cloning vector set for high-throughput functional analysis of genes in plants. *Plant Physiol.* **133**, 462–469 (2003).
30. Pierce, N. W. *et al.* Cand1 promotes assembly of new SCF complexes through dynamic exchange of F box proteins. *Cell* **153**, 206–215 (2013).
31. Harper, S. & Speicher, D. W. Expression and purification of GST fusion proteins. *Curr. Protoc. Protein Sci.* Chapter 6, Unit 6.6 (2008).
32. Huang, X. *et al.* *Arabidopsis* FHY3 and HY5 positively mediate induction of *COP1* transcription in response to photomorphogenic UV-B light. *Plant Cell* **24**, 4590–4606 (2012).
33. Carlsbecker, A. *et al.* Cell signalling by microRNA165/6 directs gene dose-dependent root cell fate. *Nature* **465**, 316–321 (2010).
34. Wang, L., Mai, Y. X., Zhang, Y. C., Luo, Q. & Yang, H. Q. MicroRNA171c-targeted *SCL6-II*, *SCL6-III*, and *SCL6-IV* genes regulate shoot branching in *Arabidopsis*. *Mol. Plant* **3**, 794–806 (2010).
35. Barrell, P. J. & Conner, A. J. Minimal T-DNA vectors suitable for agricultural deployment of transgenic plants. *Biotechniques* **41**, 708–710 (2006).
36. Heisler, M. G. *et al.* Patterns of auxin transport and gene expression during primordium development revealed by live imaging of the *Arabidopsis* inflorescence meristem. *Curr. Biol.* **15**, 1899–1911 (2005).
37. Reddy, G. V. & Meyerowitz, E. M. Stem-cell homeostasis and growth dynamics can be uncoupled in the *Arabidopsis* shoot apex. *Science* **310**, 663–667 (2005).
38. Sugimoto, K., Jiao, Y. & Meyerowitz, E. M. *Arabidopsis* regeneration from multiple tissues occurs via a root development pathway. *Dev. Cell* **18**, 463–471 (2010).
39. Roeder, A. H., Ferrandiz, C. & Yanofsky, M. F. The role of the REPLUMLESS homeodomain protein in patterning the *Arabidopsis* fruit. *Curr. Biol.* **13**, 1630–1635 (2003).
40. Bowler, C. *et al.* Chromatin techniques for plant cells. *Plant J.* **39**, 776–789 (2004).
41. Truernit, E. *et al.* High-resolution whole-mount imaging of three-dimensional tissue organization and gene expression enables the study of phloem development and structure in *Arabidopsis*. *Plant Cell* **20**, 1494–1503 (2008).



**Extended Data Figure 1 | Interaction between WUS/WOX and HAM family transcriptional regulators.** **a**, LacZ activity in yeast-two-hybrid assays testing interactions between WUS and HAM2, HAM3 or HAM4. Error bars show mean  $\pm$  s.e.m. ( $n = 3$  biological replicates). \*\* $P < 0.01$ ; \*\*\* $P < 0.001$  (two-tailed  $t$ -test, compared with DBD-WUS/AD). **b–o**, BiFC analyses in tobacco transient assays with HAM and WOX family genes. **c–q**, Tobacco was co-transformed with GFPn-WUS and GFPc-HAM3 (**b**), GFPn-WUS and GFPc-HAM4 (**c**), GFPn-WUS and GFPc-FAMA (**d**), GFPn-WUS and GFPc-BARD1 (**e**), GFPn-WOX4 and GFPc-HAM1 (**f**), GFPn-WOX4

and GFPc-FAMA (**g**), GFPn-FAMA and GFPc-HAM1 (**h**), GFPn-WOX5 and GFPc-HAM1 (**i**), GFPn-WOX5 and GFPc-HAM2 (**j**), GFPn-WOX5 and GFPc-HAM4 (**k**), GFPn-WOX5 and GFPc-FAMA (**l**), GFPn-WOX5 and GFPc-BARD1 (**m**), GFPn-BARD1 and GFPc-HAM1 (**n**), GFPn-BARD1 and GFPc-HAM2 (**o**), GFPn-BARD1 and GFPc-HAM4 (**p**), or GFPn-FAMA and GFPc-HAM4 (**q**). BARD1 and FAMA proteins are both included as negative controls. Left panel: GFP channel; middle panel: propidium iodide (PI) staining channel; right panel: merged channels. Scale bars, 20  $\mu$ m.

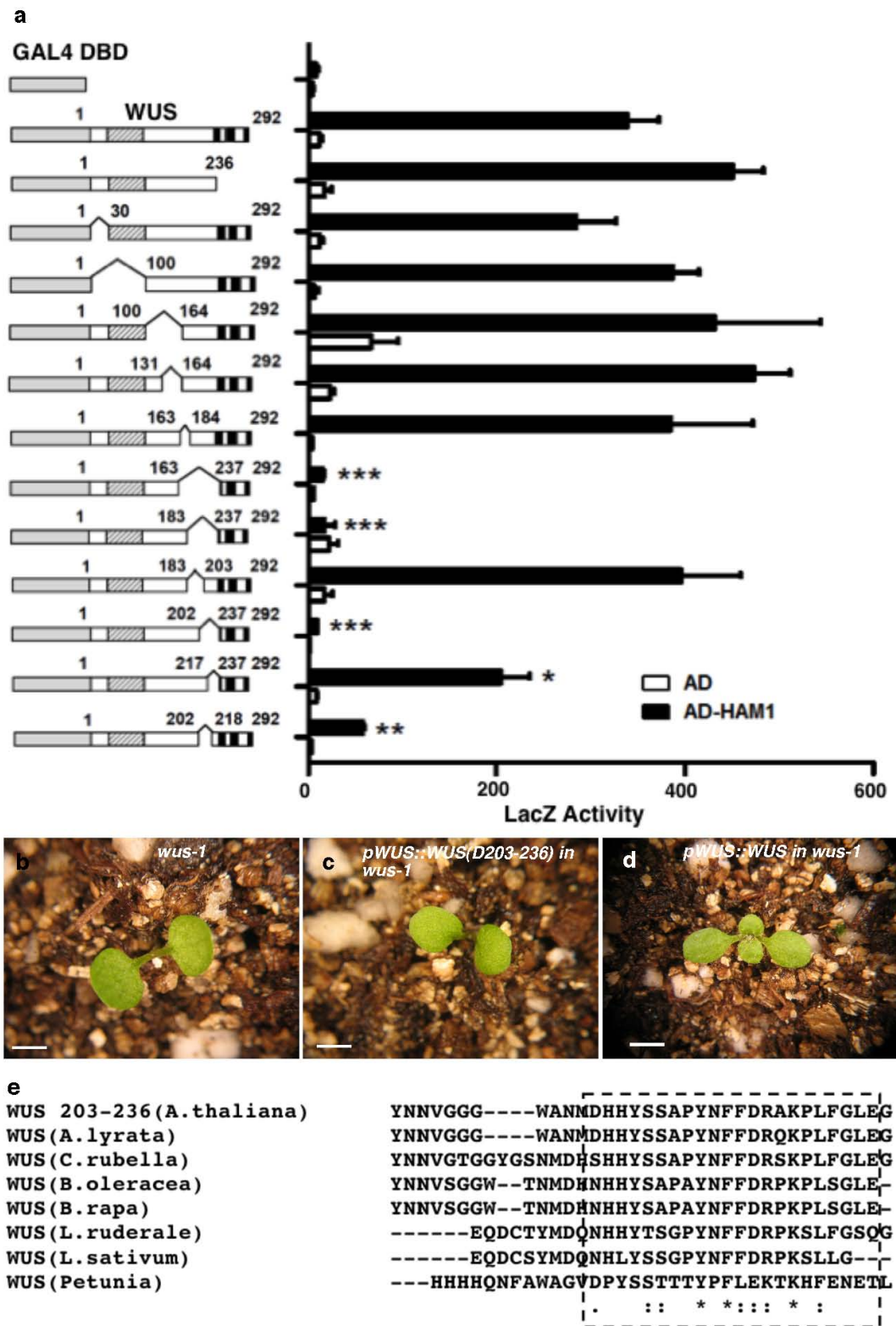




**Extended Data Figure 2 | An N-terminal region of HAM1 is important for WUS–HAM1 interaction and is essential for HAM1 function in stem cell maintenance.** **a**, Yeast two-hybrid assay of interactions between WUS and various deleted derivatives of HAM1. Deleting amino acids 117 to 230 (D117–230) from HAM1 compromised the WUS–HAM1 interaction. Left, box diagrams of the HAM1 derivatives. Shaded boxes indicate the GRAS domains. Numbers indicate amino acid residues. Error bars show mean  $\pm$  s.e.m. ( $n = 3$  biological replicates). \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  (two-tailed  $t$ -test, compared with full-length AD-HAM1). **b–g**, The complementation of the *ham1;2;4* triple mutant requires amino acids 117–230. The early termination phenotype of *ham1;2;4* (**b**, **e**) was not complemented by HAM1(D117–230) driven by a *HAM1* promoter and

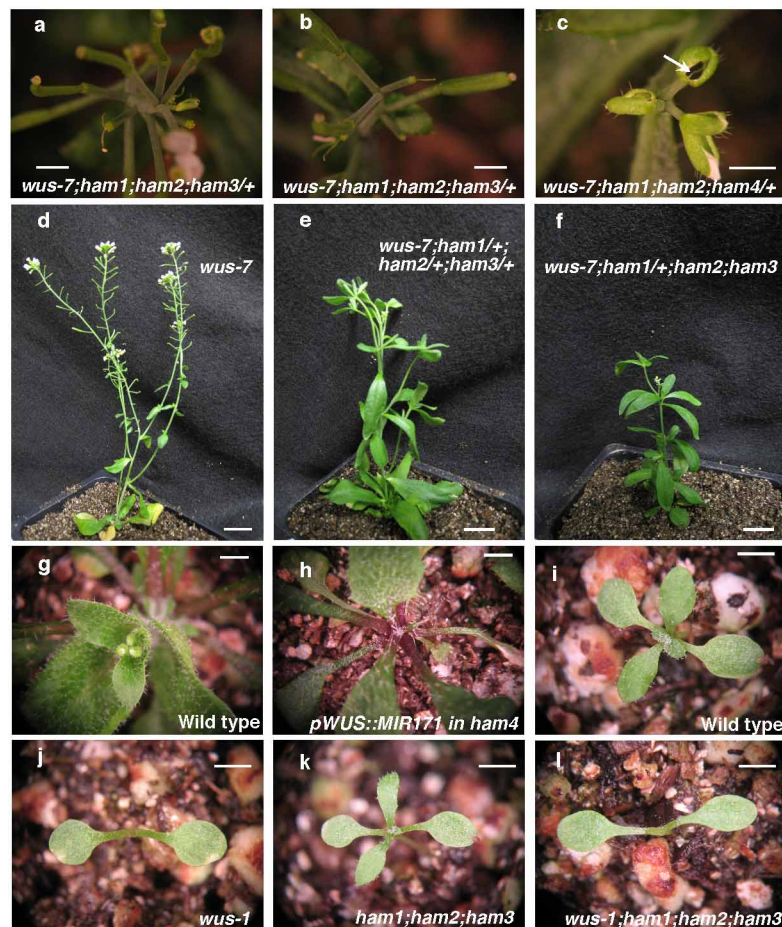
3' untranslated region (UTR) (**c**, **f**), but was fully complemented by wild-type *HAM1* (**d**, **g**). **b**, **c**, Arrows indicate the early terminated inflorescences. **h–j**, Amino acid sequence alignment of the HAM1 N-terminal domains (117–230) using Clustal Omega. **h**, Sequence alignment of the N-terminal domains among three *Arabidopsis* HAM members. **i**, Sequence alignment of partial N-terminal domains in HAM from *A. thaliana*, *A. lyrata*, *Capsella rubella*, *Brassica rapa* and *Petunia*. **j**, Sequence alignment of partial HAM1 N-terminal domains in HAM from *A. thaliana*, *A. lyrata*, *C. rubella*, *B. oleracea*, *B. rapa* and *Petunia*. Asterisks indicate amino acids that are the same; dots indicate similar amino acids. The conserved regions are boxed. Scale bars: 10 mm (**b**, **c**, **g**); 40 mm (**d**); 20 mm (**e**, **f**).





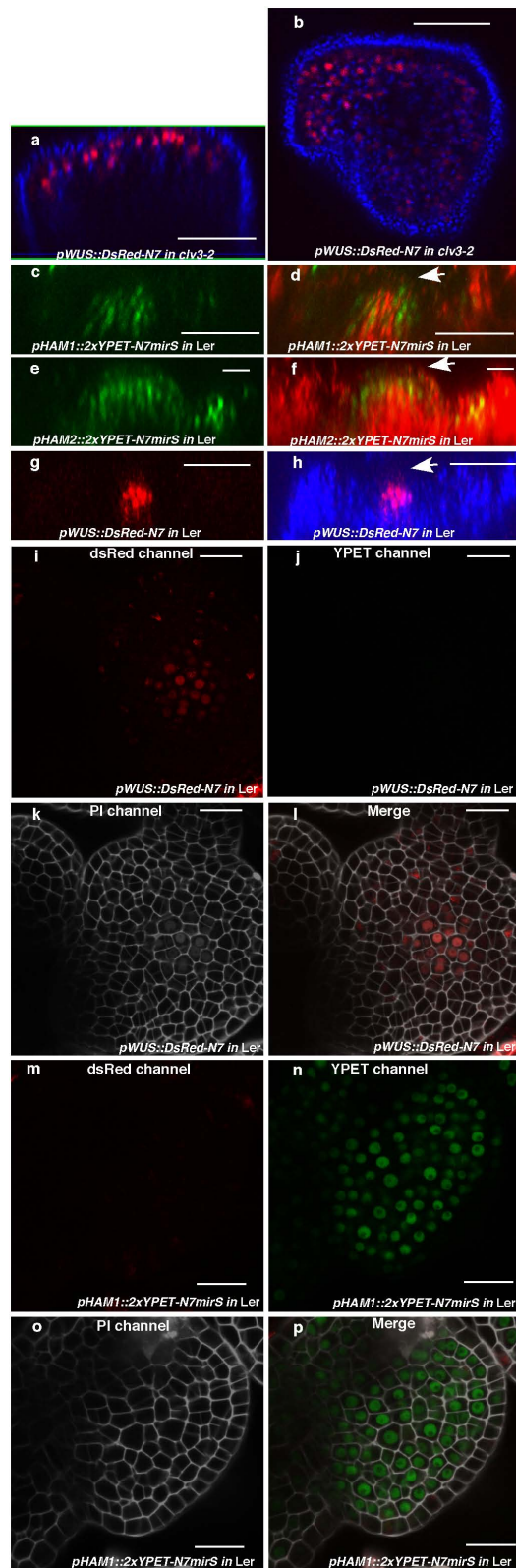
**Extended Data Figure 3 | A C-terminal region of WUS is important for WUS-HAM1 interaction and is essential for WUS function in stem cell maintenance.** **a**, Yeast-two-hybrid assay of interactions between HAM1 and various deleted WUS derivatives. Deleting amino acids 203 to 236 (D203-236) from WUS greatly compromised the WUS-HAM1 interaction. Left, box diagrams of the deleted WUS derivatives. Shaded boxes indicate the homeodomain; the three black boxes indicate the acidic domains, the WUS box and the EAR motif, respectively. Numbers indicate amino acid residues. Error bars show mean  $\pm$  s.e.m. ( $n = 3$  biological replicates). \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  (two-tailed  $t$ -test, compared with DBD-WUS full-length).

**b-d**, WUS function requires the same region that is important for WUS-HAM1 interaction. The terminated shoot meristem phenotype of *wus-1* (**b**) was not complemented by WUS(D203-236) driven by the WUS promoter and 3' UTR (**c**), and was fully complemented by the wild-type WUS (**d**). **e**, Amino acid sequence alignment of C-terminal regions of WUS from *A. thaliana*, *A. lyrata*, *C. rubella*, *B. oleracea*, *B. rapa*, *Lepidium ruderales*, *L. sativum* and *Petunia*, using Clustal Omega. Asterisks indicate amino acids that are the same; dots indicate similar amino acids. The conserved regions are boxed. Scale bars, 2 mm (**b-d**).



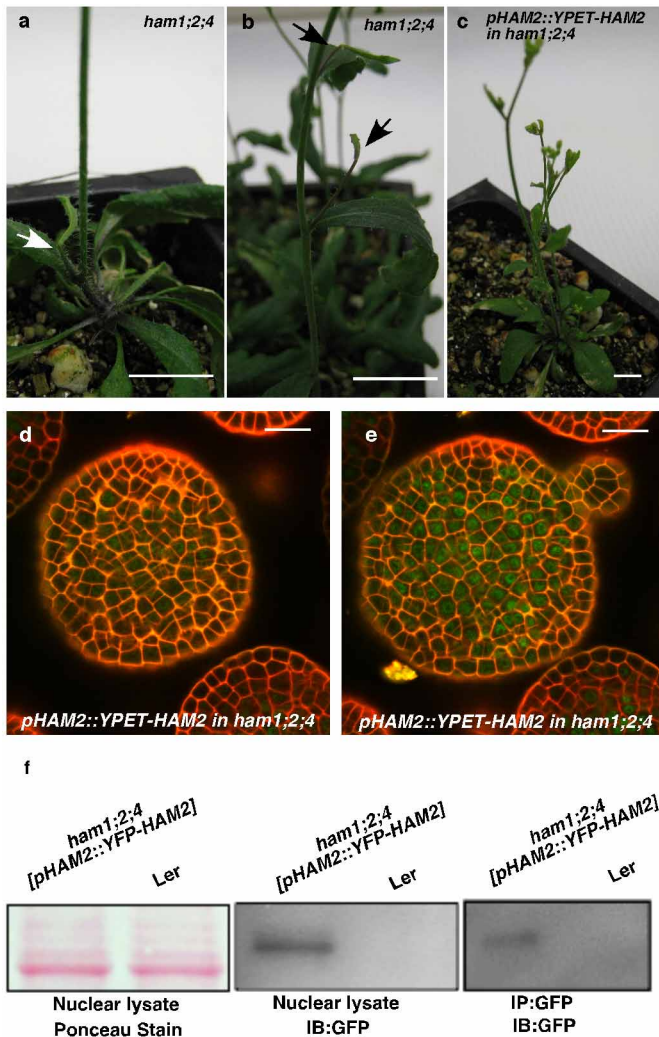
**Extended Data Figure 4 | Genetic interaction between WUS and HAM family members.** **a, b,** The secondary inflorescence meristems initiated from axillary meristems in *wus-7;ham1;2* homozygotes with *ham3/+* terminate prematurely. **c,** *wus-7;ham1;2* homozygotes with *ham4/+* display early termination of the main inflorescence meristem and lack of carpels in flowers (indicated by arrow). **d–f,** WUS and HAM family members interact genetically in a dose-dependent manner. *wus-7* (**d**) formed functional shoot apices and normal stature, but *wus-7;ham1/+;ham2/+;ham3/+* (**e**) enhanced the *wus-7* phenotype, and *wus-7;ham1/+;ham2;ham3* (**f**) showed stronger enhancement, with reduced flower numbers and plant stature, and an

elongated vegetative stage, resembling a *wus* strong allele. Plants are at 36 days after germination. **g, h,** Downregulation of *HAM1*, *HAM2* and *HAM3* in *ham4* shoot meristems leads to an early termination phenotype. Compared to wild type (Col) (**g**), *pWUS::MIR171* in *ham4* (**h**) showed terminated vegetative meristems. **i–l,** WUS is required for the functions of *HAM1*, *HAM2* and *HAM3*. At 11 days after germination, compared with Ler wild type (**i**) and *ham1;2;3* (**k**), which formed functional vegetative meristem and leaf primordia, *wus-1;ham1;2;3* (**l**) displays terminated vegetative meristems similar to *wus-1* (**j**). Scale bars, 2 mm.



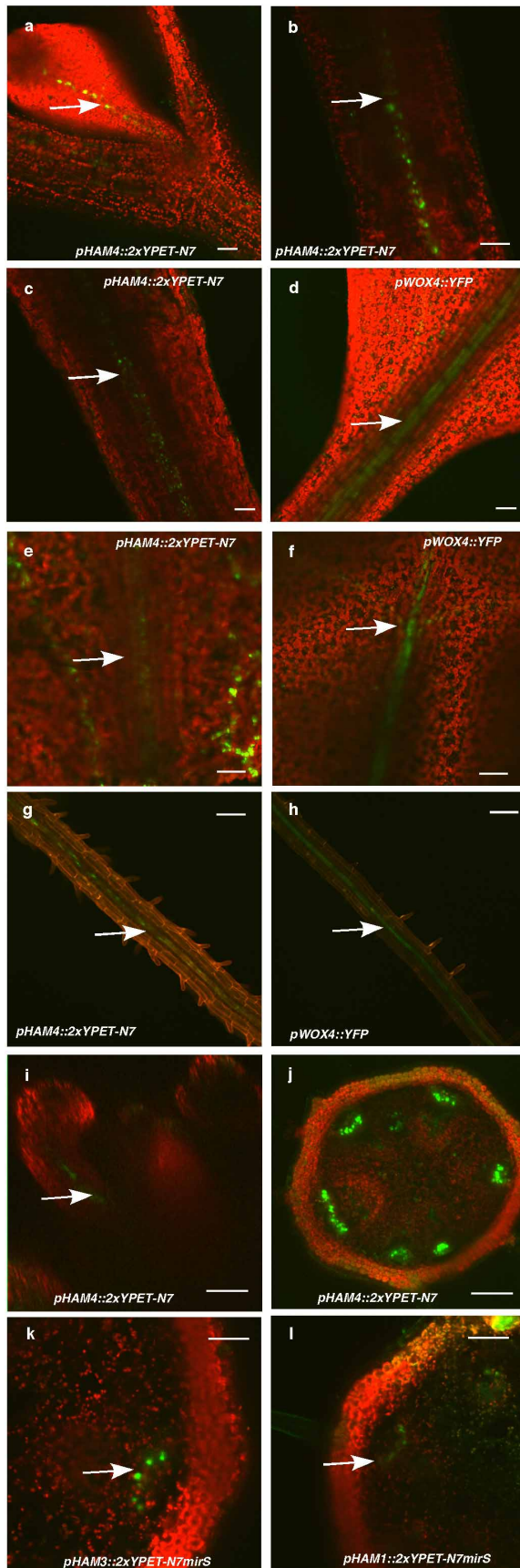
**Extended Data Figure 5 | Expression of *HAM1*, *HAM2* and *WUS* in the SAMs.** **a, b,** *WUS* expression in *clv3-2*. Orthogonal (**a**) and top (**b**) views of *pWUS::DsRed-N7* expression (red) and chlorophyll autofluorescence (blue) in the same *clv3-2* inflorescence meristem. **c–h,** Comparison between expression patterns of *HAM1*, *HAM2* and *WUS* in vegetative meristems. **c,** Orthogonal view of *pHAM1::2xYPET-N7mirS* expression (green) in *Ler* vegetative meristem. **d,** Orthogonal view of *pHAM1::2xYPET-N7mirS* expression (green) together with chlorophyll autofluorescence (red) in the same vegetative meristem shown in **c**, indicating that *HAM1* is expressed in the rib meristem. **e,** Orthogonal view of *pHAM2::2xYPET-N7mirS* expression (green) in *Ler* vegetative meristem. **f,** Orthogonal view of *pHAM2::2xYPET-N7mirS* expression (green) together with chlorophyll autofluorescence (red) in the same vegetative meristem shown in **e**, indicating that *HAM2* is highly expressed in the rib meristem. **g,** Orthogonal view of *pWUS::DsRed-N7* expression (red) in *Ler* vegetative meristem. **h,** Orthogonal view of *pWUS::DsRed-N7* expression (red) together with chlorophyll autofluorescence (blue) in the same vegetative meristem shown in **g**, indicating that *WUS* is expressed in the rib meristem. Arrows indicate the positions of the L1 cell layer. **i–p,** Control images confirming the specificity of confocal spectral settings for Fig. 3 (**e–l**). The SAMs from the *pWUS::DsRed-N7* line (**i–l**) or *pHAM1::2xYPET-N7mirS* line (**m–p**) were imaged from the same three separated channels used in Fig. 3 (**e–l**). There is no spectral bleed-through of YPET signal into the dsRed channel (**m**), nor dsRed signal into the YPET channel (**j**). **i, m,** dsRed channel (red); **j, n,** YPET channel (green); **k, o,** PI staining channel (grey); **l, p,** merged all three channels. Scale bars: 50  $\mu$ m (**a–d, g, h**); 20  $\mu$ m (**e, f, i–p**).



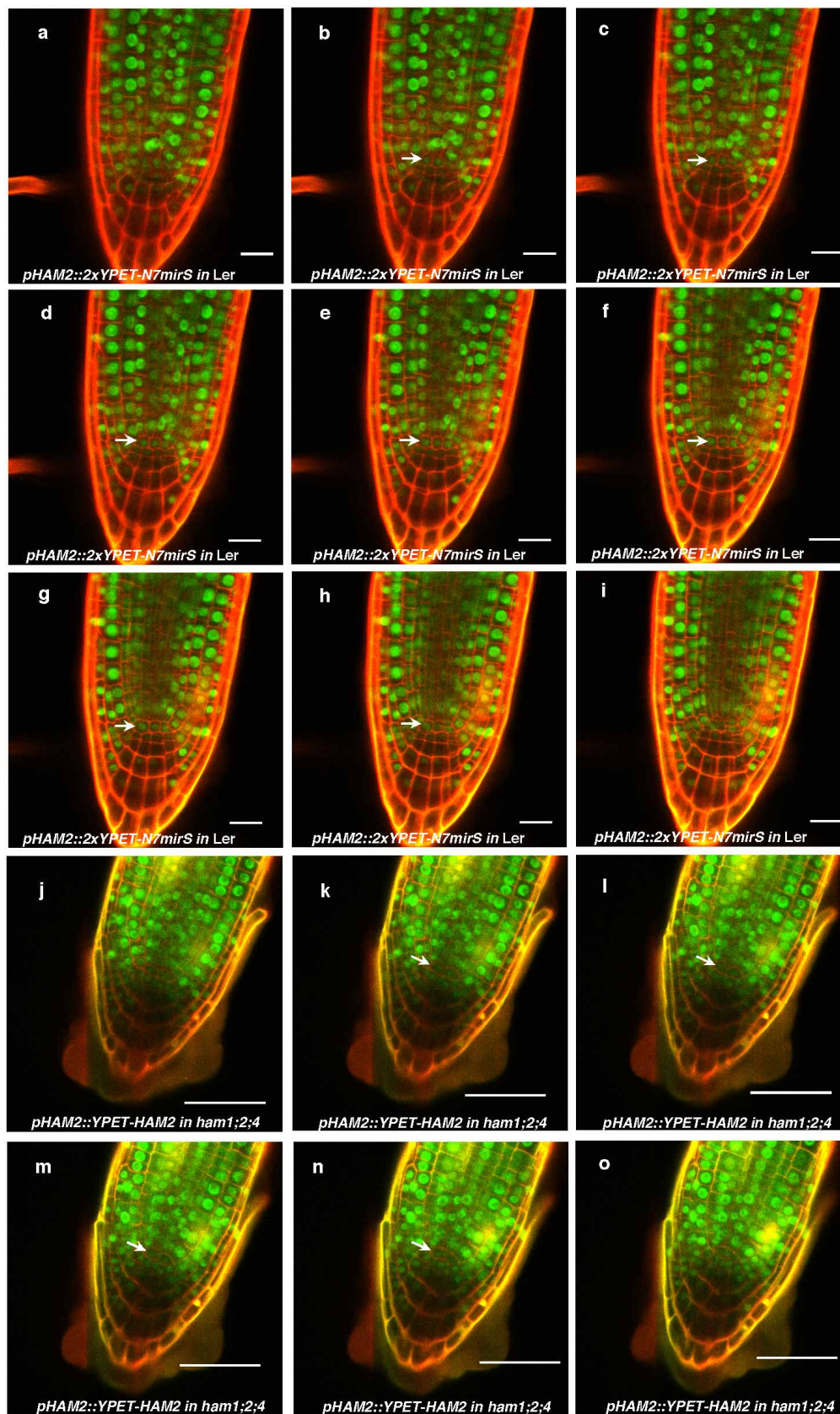


**Extended Data Figure 6 | *pHAM2::YFP-HAM2* (*pHAM2::YPET-HAM2*) complemented the *ham1;2;4* mutant and was expressed in the centre of SAMs. a–c**, The early termination phenotype of *ham1;2;4* (a, b) was completely complemented by YPET-HAM2 driven by the *HAM2* promoter and 3' UTR (c), indicating that the promoter used for *HAM2* transcriptional and translational reporters is functional and that the fusion protein (YPET-HAM2) is also functional *in vivo*. a, b, Arrows indicate early terminated apices. a–c, Scale bars, 10 mm. d, e, Different Z sections from the same SAM from a *ham1;2;4* [pHAM2::YPET-HAM2] plant depicted in Fig. 3m, n shows expression of *pHAM2::YPET-HAM2* translational marker (green) in L2 (d) and L3 (e), together with PI as counter stain (red). d, e, Scale bars, 20 μm. f, Immunoblot with anti-GFP antibody validates the presence of YFP-HAM2 (YPET-HAM2) in both nuclear lysate and nuclear proteins immunoprecipitated with GFP-Trap from *ham1;2;4* [pHAM2::YFP-HAM2] line used in ChIP experiment (Fig. 2n, o). IB, immunoblot; IP, immunoprecipitation.





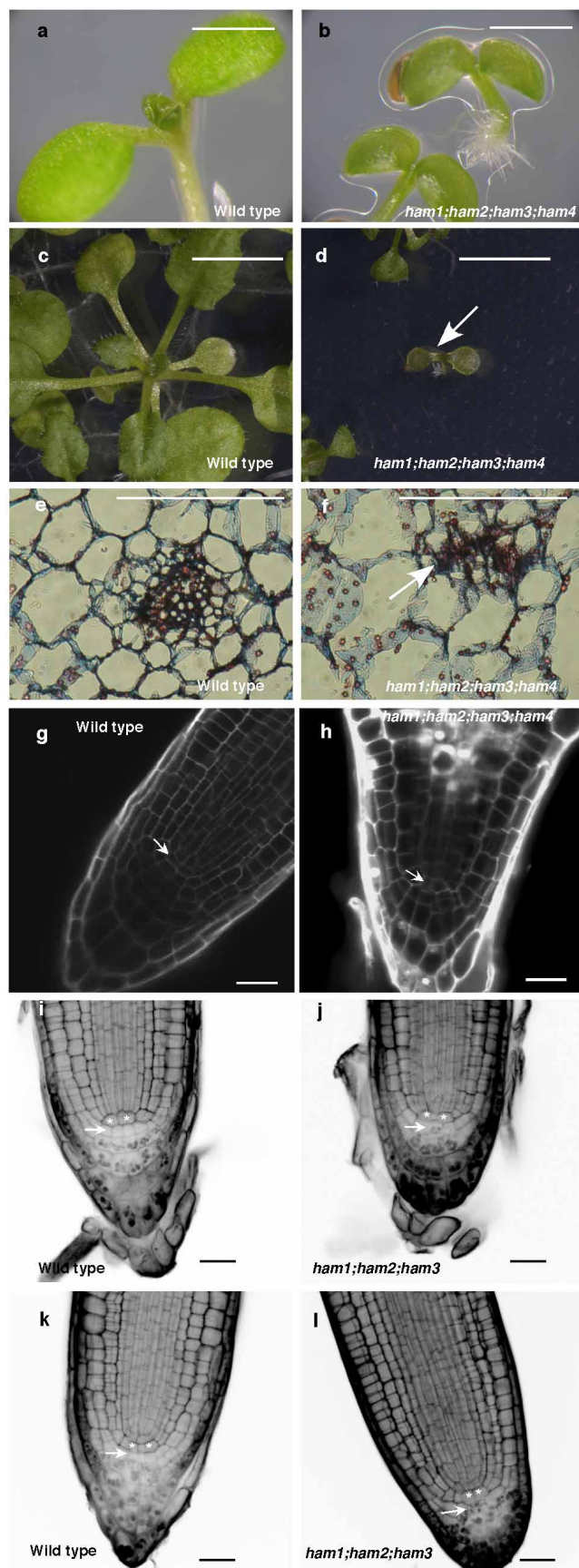
**Extended Data Figure 7 | Expression patterns of *HAM* genes in comparison with *WOX4*.** **a**, *pHAM4::2xYPET-N7* (green, indicated by arrow) is expressed in procambium cells of the first leaf. **b**, *pHAM4::2xYPET-N7* (green, indicated by arrow) is expressed in vasculature in the 7-day-old hypocotyl. **c–h**, Comparison of *pHAM4::2xYPET-N7* (green, indicated by arrow) and *pWOX4::YFP* (green, arrow indicated) expression patterns in vasculature cells in the 7-day-old leaf petiole (**c**, **d**), 20-day-old leaf petiole (**e**, **f**) and 7-day-old root (**g**, **h**). **i**, Orthogonal view of *pHAM4::2xYPET-N7* (green, indicated by arrow) expression in flower vasculature. **j**, Procambium-specific expression of *pHAM4::2xYPET-N7* in stems from 1-cm bolting plants. **k–l**, Procambium-specific expression of *pHAM3::2xYPET-N7mirS* (**k**) and *pHAM1::2xYPET-N7mirS* (**l**) in transverse sections of stems from 1-cm bolting plants. Red represents chlorophyll autofluorescence (**a–f**, **i–l**), or PI staining (**g**, **h**). Scale bars: 50  $\mu\text{m}$  (**a**, **h–i**, **k–l**); 100  $\mu\text{m}$  (**b–g**, **j**).



**Extended Data Figure 8 | Expression patterns of *HAM2* transcriptional and translational reporters in root meristems.** a–i, Complete stacks of confocal sections through the root tip demonstrate that *pHAM2::2xYPET-N7mirS* (green) is expressed in the quiescent centre cells (indicated by arrow) and in cells above the quiescent centre within the root meristem. j–o, Expression patterns of *HAM2* translational reporters in *ham1;2;4* root meristems.

Complete stacks of confocal sections through the root tip demonstrate that *pHAM2::YPET-HAM2* (green) is present in the quiescent centre cells (indicated by arrows) and the cells above the quiescent centre within the root meristem in the *ham1;2;4* mutant. Cellular outlines were stained with PI (red). Scale bars: 20  $\mu\text{m}$  (a–i); 50  $\mu\text{m}$  (j–o).





### Extended Data Figure 9 | HAM family regulates various stem cell niches.

**a–d**, Growth arrest of *ham1;2;3;4* at the seedling stage. **a**, **b**, Imaging of Ler wild-type (**a**) and homozygous *ham1;2;3;4* (**b**) seedlings at 7 DAG. **c**, **d**, Imaging of wild-type (**c**) and homozygous *ham1;2;3;4* (**d**) (indicated by arrow) seedlings at 26 DAG. **e**, **f**, Transverse section of leaves from wild-type (**e**) and *ham1;2;3;4* (**f**) at 7 DAG. **f**, Arrow indicates undifferentiated/undetermined cell mass. **g**, **h**, Confocal imaging of root meristem from wild-type (**g**) and *ham1;2;3;4* (**h**) seedlings at 7 DAG. *ham1;2;3;4* displayed enlarged cells with abnormal shapes at the quiescent centre (indicated by arrows) and CSC positions. **g**, **h**, Cellular outlines were visualized with PI staining (white). **i–l**, mPS-PI<sup>11</sup> stains indicate that HAM genes regulate root cell differentiation. Some CSCs (arrow indicated) undergo differentiation with starch accumulated and stained in homozygous *ham1;2;3* (**j**, **l**), but none of them can be stained in Ler wild type (**i**, **k**). Asterisks mark the quiescent centre cells. Scale bars: 5 mm (**c**, **d**); 1 mm (**a**, **b**, **e**, **f**); 20 μm (**g–l**).



**Extended Data Figure 10 | Interaction between WOX and HAM homologues from tomato (*Solanum lycopersicum*).** a, b, f, BiFC analyses in tobacco transient assays demonstrated that tomato WUS (NCBI gene accession number 543793) physically interacted with a putative tomato HAM homologue (sequence accession number: LEFL2052P11 from Kazusa Full-length Tomato cDNA database) (a) identified based on its sequence homology to HAM from *Arabidopsis* and *Petunia* (f), and that tomato WOX4

(ref. 10) (NCBI gene accession number 100301933) physically interacted with the putative tomato HAM homologue (b). c–e, BARD1 protein is included as a negative control. Left panel: GFP channel; middle panel: PI staining channel; right panel: merged channels. Scale bars, 20  $\mu$ m. f, Amino acid sequence alignment of a putative tomato HAM, *Arabidopsis* HAM1 and *Petunia* HAM using Clustal Omega. Asterisks indicate amino acids that are the same; dots indicate similar amino acids.



# Broad CTL response is required to clear latent HIV-1 due to dominance of escape mutations

Kai Deng<sup>1</sup>, Mihaela Perteau<sup>1,2</sup>, Anthony Rongvaux<sup>3</sup>, Leyao Wang<sup>4</sup>, Christine M. Durand<sup>1</sup>, Gabriel Ghiaur<sup>5</sup>, Jun Lai<sup>1</sup>, Holly L. McHugh<sup>1</sup>, Haiping Hao<sup>6</sup>, Hao Zhang<sup>7</sup>, Joseph B. Margolick<sup>7</sup>, Cagan Gurer<sup>8</sup>, Andrew J. Murphy<sup>8</sup>, David M. Valenzuela<sup>8</sup>, George D. Yancopoulos<sup>8</sup>, Steven G. Deeks<sup>9</sup>, Till Strowig<sup>3</sup>, Priti Kumar<sup>10</sup>, Janet D. Siliciano<sup>1</sup>, Steven L. Salzberg<sup>2,11</sup>, Richard A. Flavell<sup>3,12</sup>, Liang Shan<sup>3</sup> & Robert F. Siliciano<sup>1,13</sup>

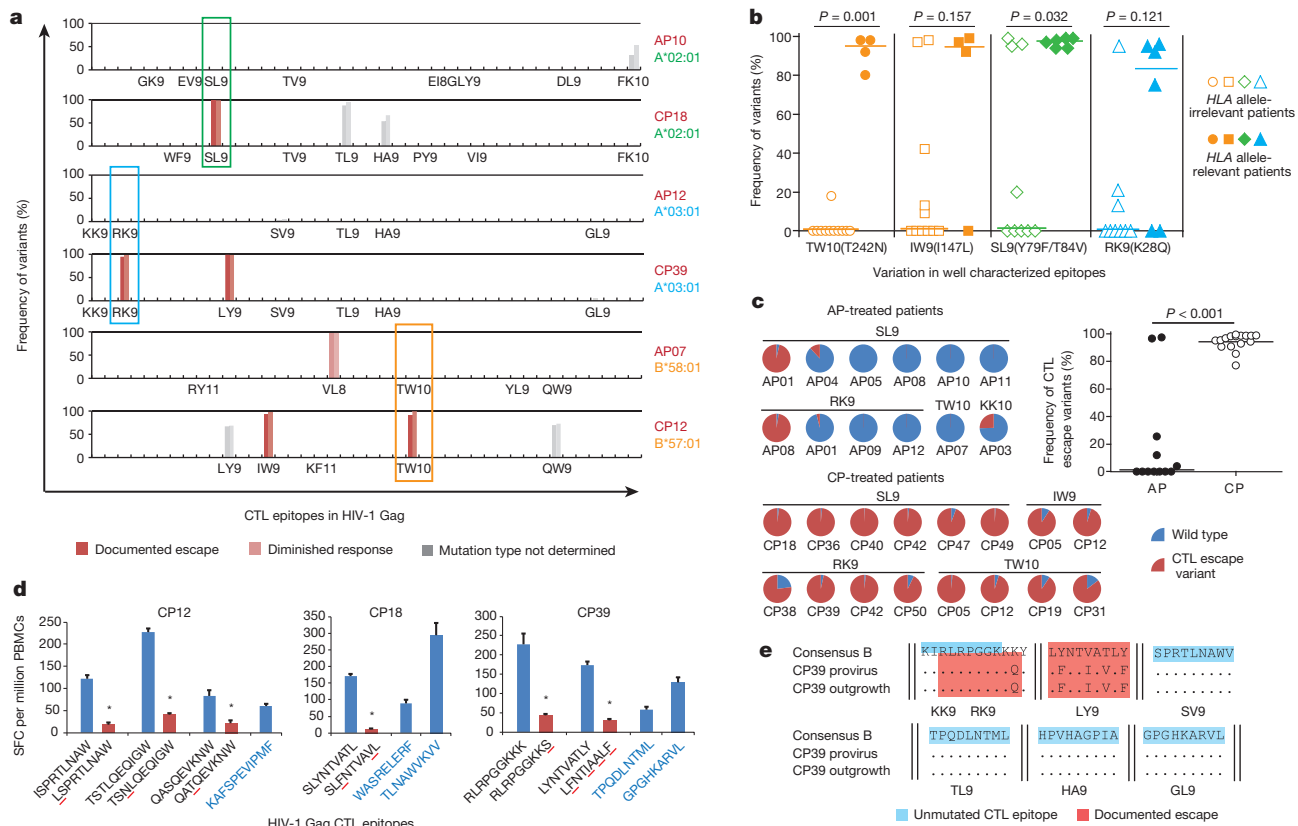
Despite antiretroviral therapy (ART), human immunodeficiency virus (HIV)-1 persists in a stable latent reservoir<sup>1,2</sup>, primarily in resting memory CD4<sup>+</sup> T cells<sup>3,4</sup>. This reservoir presents a major barrier to the cure of HIV-1 infection. To purge the reservoir, pharmacological reactivation of latent HIV-1 has been proposed<sup>5</sup> and tested both *in vitro* and *in vivo*<sup>6–8</sup>. A key remaining question is whether virus-specific immune mechanisms, including cytotoxic T lymphocytes (CTLs), can clear infected cells in ART-treated patients after latency is reversed. Here we show that there is a striking all or none pattern for CTL escape mutations in HIV-1 Gag epitopes. Unless ART is started early, the vast majority (>98%) of latent viruses carry CTL escape mutations that render infected cells insensitive to CTLs directed at common epitopes. To solve this problem, we identified CTLs that could recognize epitopes from latent HIV-1 that were unmutated in every chronically infected patient tested. Upon stimulation, these CTLs eliminated target cells infected with autologous virus derived from the latent reservoir, both *in vitro* and in patient-derived humanized mice. The predominance of CTL-resistant viruses in the latent reservoir poses a major challenge to viral eradication. Our results demonstrate that chronically infected patients retain a broad-spectrum viral-specific CTL response and that appropriate boosting of this response may be required for the elimination of the latent reservoir.

HIV-1 establishes latent infection in resting CD4<sup>+</sup> T cells<sup>3,4</sup>. Recent efforts to eradicate HIV-1 infection have focused on reversing latency without global T-cell activation<sup>5</sup>. However, inducing HIV-1 gene expression in latently infected cells is not sufficient to cause the death of these cells if they remain in a resting state<sup>9</sup>. Boosting HIV-1-specific immune responses, including CTL responses, may be required for clearance of the latent reservoir<sup>9</sup>. CTLs have a significant role in suppressing HIV-1 replication in acute infection<sup>10–14</sup>. Because of this strong selective pressure, HIV-1 quickly acquires mutations to evade CTL recognition<sup>12,13,15–18</sup>. CTL escape has been studied primarily through the analysis of plasma virus<sup>12,13,16,18–20</sup>, and CTL-based vaccines have been designed based on conserved epitopes<sup>21,22</sup>. A systematic investigation of CTL escape in the latent reservoir will be of great importance to the ongoing CTL-based virus eradication efforts, because latent HIV-1 probably represents the major source of viral rebound after treatment interruption. Earlier studies have suggested the presence of CTL escape mutations in proviral DNA<sup>15,17</sup>, but it remains unclear to what extent the latent reservoir in resting CD4<sup>+</sup> T cells is affected by CTL escape, whether mutations detected in proviral DNA are representative of the very small fraction of proviruses that are replication competent, and, most importantly, whether the CTL response can recognize and clear infected cells after latency is reversed.

To investigate CTL escape variants in the latent reservoir, we deep sequenced the proviral HIV-1 DNA in resting CD4<sup>+</sup> T cells from 25 patients (Extended Data Table 1). Among them, 10 initiated ART during the acute phase (AP; within 3 months of infection) while the other 15 initiated ART during the chronic phase (CP) of infection. The sequencing was focused on Gag because it is an important target of the CTL response<sup>23</sup> and is highly conserved, which facilitates the detection of escape variants. Our data show that previously documented CTL escape variants completely dominate the viral reservoirs of nearly all CP-treated patients (Extended Data Fig. 1 and Supplementary Table 1). This trend is especially obvious for several well characterized CTL epitopes: the human leukocyte antigen (HLA)-A2-restricted epitope SLYNTVATL (SL9), the HLA-A3-restricted epitope RLRPGGKKK (RK9) and the HLA-B57/58-restricted epitope TSTLQEQIGW (TW10) (Fig. 1a and Extended Data Fig. 1, highlighted in coloured boxes.). In these epitopes, close to 100% of the sequences harboured escape mutations. Comparison of mutation frequencies between HLA allele-relevant and -irrelevant epitopes in CP-treated patients suggests that the CTL escape mutations identified are specific to each patient's HLA type (Fig. 1b). By contrast, except for SL9 from patient AP01 and RK9 from patient AP08, few if any CTL escape mutations were archived in AP-treated patients (Fig. 1c and Extended Data Fig. 1). The striking difference between AP- and CP-treated patients (Fig. 1c) indicates that, unless treatment is initiated within the first several months of infection, the latent reservoir becomes almost completely dominated by variants resistant to dominant CTL responses.

To confirm that variants detected at high frequency in the latent reservoir represent functional CTL escape mutants, cells from seven CP-treated subjects were tested for reactivity to synthetic peptides representing wild-type and mutant versions of the relevant epitopes. As expected, there were only minimal responses to previously documented CTL escape mutants by patient CD8<sup>+</sup> T cells, and no *de novo* response was detected (Fig. 1d and Extended Data Fig. 2). In contrast, all tested subjects retained a strong response to peptides representing the wild-type epitopes, suggesting that the wild-type virus was initially transmitted, with subsequent evolution of CTL escape variants. Most HIV-1 proviruses detected in patients are defective<sup>24</sup>. Therefore, to determine whether these CTL escape variants can be reactivated and lead to viral rebound if therapy is stopped, we isolated replication-competent viruses from the latent reservoirs of nine CP-treated patients. We found that all the dominant CTL escape mutations that had been identified in proviruses in resting CD4<sup>+</sup> T cells were also present in the replication-competent viruses that grew out after T-cell activation (Fig. 1e and Extended Data Fig. 3), indicating that these CTL escape variants not only dominate the population of proviruses, but can also be released and replicate once latency is reversed.

<sup>1</sup>Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. <sup>2</sup>Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. <sup>3</sup>Department of Immunobiology, Yale University School of Medicine, New Haven, Connecticut 06510, USA. <sup>4</sup>Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut 06510, USA. <sup>5</sup>Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. <sup>6</sup>Deep Sequencing and Microarray Core, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. <sup>7</sup>Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA. <sup>8</sup>Regeneron Pharmaceuticals Inc., Tarrytown, New York 10591, USA. <sup>9</sup>Department of Medicine, University of California, San Francisco, San Francisco, California 94110, USA. <sup>10</sup>Department of Medicine, Yale University School of Medicine, New Haven, Connecticut 06510, USA. <sup>11</sup>Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. <sup>12</sup>Howard Hughes Medical Institute, New Haven, Connecticut 06510, USA. <sup>13</sup>Howard Hughes Medical Institute, Baltimore, Maryland 21205, USA.



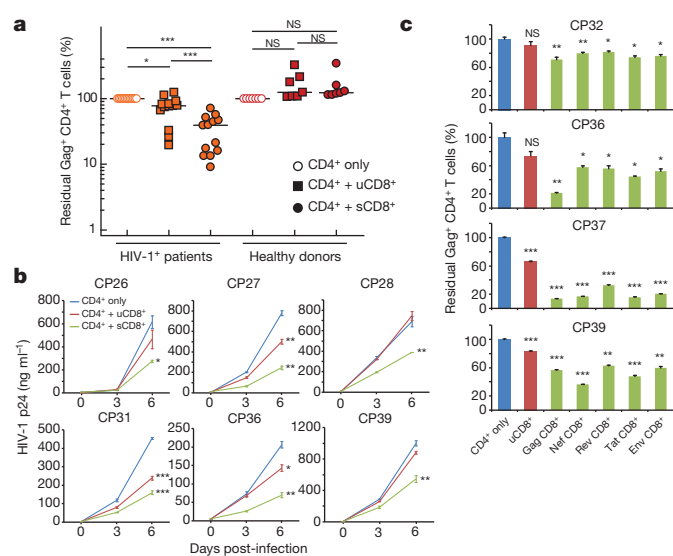
**Figure 1 | CTL escape variants dominate the latent reservoir of CP-treated but not AP-treated patients.** **a**, Frequency of variants in Gag CTL epitopes in proviruses from resting CD4<sup>+</sup> T cells. Representative results from six patients are shown. Only optimal CTL epitopes relevant to each patient's HLA type are listed. Results from both PacBio (left bar) and MiSeq (right bar) sequencing are shown. The effect on CTL recognition (denoted by colour) is determined from information in the Los Alamos National Laboratory (LANL) HIV Molecular Immunology Database. **b**, CTL escape variants identified by sequencing are specific to HLA type. Frequencies of documented escape-associated variants in four well characterized epitopes are shown for all 15 CP-treated patients. Medians and *P* values from Mann–Whitney test are shown. **c**, Comparison of CTL escape variant frequency in proviruses between CP- and AP-treated patients. Only well characterized epitopes are shown.

We next asked whether the host CTL response could recognize and eliminate the cells infected with these escape variants. We infected activated CD4<sup>+</sup> T cells from these patients with autologous, replication-competent virus derived from the latent reservoir (Extended Data Fig. 4a). The infected cells were then co-cultured with autologous CD8<sup>+</sup> T cells, either unstimulated or pre-stimulated, to assess HIV-1-specific cytolytic activity. Non-specific activation of CD8<sup>+</sup> T cells was not observed after co-culture with phytohaemagglutinin (PHA)-activated CD4<sup>+</sup> T cells (Extended Data Fig. 4b). From all 13 CP-treated subjects tested, CD8<sup>+</sup> T cells pre-stimulated by a Gag peptide mixture efficiently killed autologous infected CD4<sup>+</sup> T cells (median 61% elimination), while unstimulated CD8<sup>+</sup> T cells from most subjects had significantly less effect (median 23% elimination) (Fig. 2a and Extended Data Fig. 4c, d). CD8<sup>+</sup> T cells from 7/7 healthy donors completely failed to eliminate autologous infected cells (Fig. 2a), confirming that the observed killing was HIV-1 specific. The killing effect was enhanced by increasing the effector to target ratio (Extended Data Fig. 5a), and was cell–cell contact dependent (Extended Data Fig. 5b). When the co-culture was maintained over time in the absence of ART, viral replication was significantly reduced, but not completely inhibited by pre-stimulated CD8<sup>+</sup> T cells (Fig. 2b). We found that peptide mixtures from other HIV-1 proteins (Nef, Tat, Rev and Env) could also boost CTL responses and facilitate the elimination of infected cells (Fig. 2c), and that CTLs pre-stimulated

Medians and *P* values from Mann–Whitney test are shown. **d**, Characterization of CTL responses against HIV-1 Gag epitopes by interferon- $\gamma$  ELISpot. The peptides tested are listed below the x-axis (black type, epitopes in which sequence variation was detected; blue type, no variation). The observed mutation is underlined in red, and CTL escape (defined by the absence of positive response after mutation) is denoted by an asterisk above the bar. The peptide concentration was 10  $\mu\text{g ml}^{-1}$ . PBMCs, peripheral blood mononuclear cells; SFC, spot-forming cell. Error bars represent standard error of the mean (s.e.m.), *n* = 3. **e**, Sequences in Gag CTL epitopes for proviral DNA and outgrowth virus from resting CD4<sup>+</sup> T cells in patient CP39. CTL epitopes with no observed variation are highlighted in blue. Epitopes with documented escape mutations are shaded in red.

by Gag peptides generally had the highest activity. Together, these results demonstrate that chronically infected patients retain CTL clones that can recognize and eliminate autologous infected CD4<sup>+</sup> T cells, despite the presence of CTL escape mutations in dominant epitopes. However, these clones require stimulation with antigen for optimal activity.

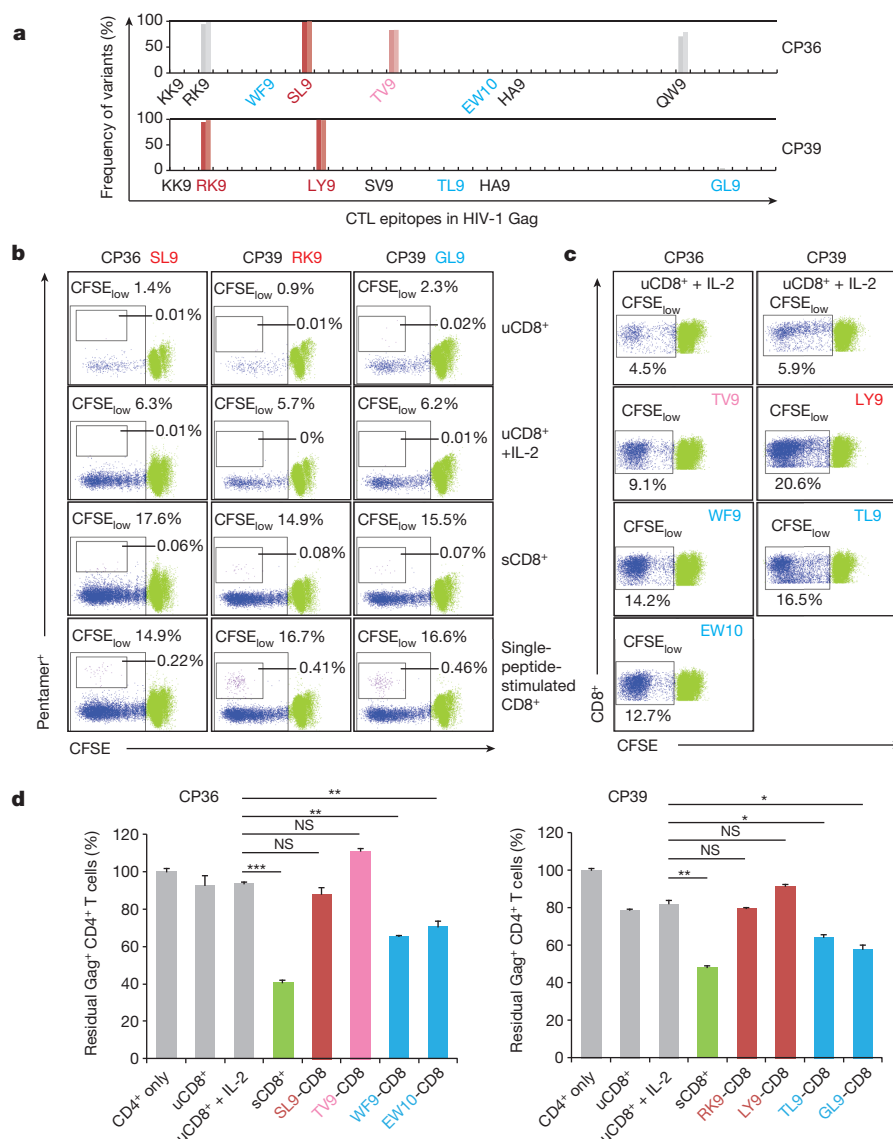
To characterize further which CTL population contributed to the elimination of cells infected by CTL escape variants, we compared the killing activity of two specific CTL populations: the population that targets epitopes in which escape has been identified and the one that targets unmutated epitopes (Fig. 3a). CD8<sup>+</sup> T cells from patients CP36 and CP39 were pre-stimulated with interleukin (IL)-2 and different synthetic peptides representing the wild-type forms of the relevant epitopes. After incubation for 6 days, each CTL population exhibited significant proliferation compared with no treatment or IL-2 alone (Fig. 3b, c). Pentamer staining for three available epitopes revealed that the number of epitope-specific CD8<sup>+</sup> T cells increased dramatically after stimulation with wild-type peptides (Fig. 3b). After co-culture with autologous target cells infected with latent reservoir-derived viruses, CTLs targeting unmutated epitopes clearly showed stronger cytolytic activity than the IL-2 only controls, while CTLs targeting epitopes with identified escaped mutations showed no significant killing (Fig. 3d). CTLs pre-stimulated by the Gag peptide mixture exhibited stronger killing than all single-peptide-stimulated populations (Fig. 3d).



**Figure 2** | CD8<sup>+</sup> T cells pre-stimulated with a mixture of Gag peptides eliminate autologous CD4<sup>+</sup> T cells infected with autologous HIV-1 from resting CD4<sup>+</sup> T cells. **a**, Pre-stimulated CD8<sup>+</sup> T cells (sCD8<sup>+</sup>) eliminate autologous infected CD4<sup>+</sup> T cells more efficiently than unstimulated CD8<sup>+</sup> T cells (uCD8<sup>+</sup>). Each symbol represents the mean of three replicates. Medians and *P* values from Mann–Whitney test are shown. **b**, sCD8<sup>+</sup> cells inhibit viral growth in autologous infected CD4<sup>+</sup> T cells with higher efficacy than uCD8<sup>+</sup> cells. p24, HIV-1 capsid (core) protein. **c**, sCD8<sup>+</sup> cells pre-stimulated by different viral peptides eliminate autologous CD4<sup>+</sup> T cells infected with viruses derived from resting CD4<sup>+</sup> T cells. **b**, **c**, Results were compared with CD4<sup>+</sup> only using paired *t*-tests. Error bars represent s.e.m., *n* = 3. \**P* < 0.05, \*\**P* < 0.01, \*\*\**P* < 0.001, NS, not significant (*P* > 0.05).

To test whether CTLs that recognize unmutated viral epitopes can inhibit HIV-1 replication and clear infected cells *in vivo*, we generated patient-derived humanized mice using an improved version of a recently

reported mouse system named MISTRG<sup>25</sup>. Whereas the previously reported MISTRG mice bear a bacterial artificial chromosome (BAC) transgene encoding human SIRP- $\alpha$ , the newly generated MIS<sup>(KI)</sup>TRG mice harbour a knock-in replacement of the endogenous mouse *Sirpa* gene with a humanized version. With humanization by knock-in replacement of the *Csf1*, *Csf2*, *Il3*, *Tpo* and *Sirpa* genes in the *Rag2*<sup>-/-</sup> *Il2rg*<sup>-/-</sup> genetic background, MIS<sup>(KI)</sup>TRG mice are highly permissive for human haematopoiesis and support the reconstitution of robust human lymphoid and myelomonocytic systems. With the demonstrated development of functional T lymphocytes and monocytes/macrophages, MIS<sup>(KI)</sup>TRG mice provide a useful humanized mouse host for HIV-1 infection studies. Bone marrow biopsies were obtained from study participants and



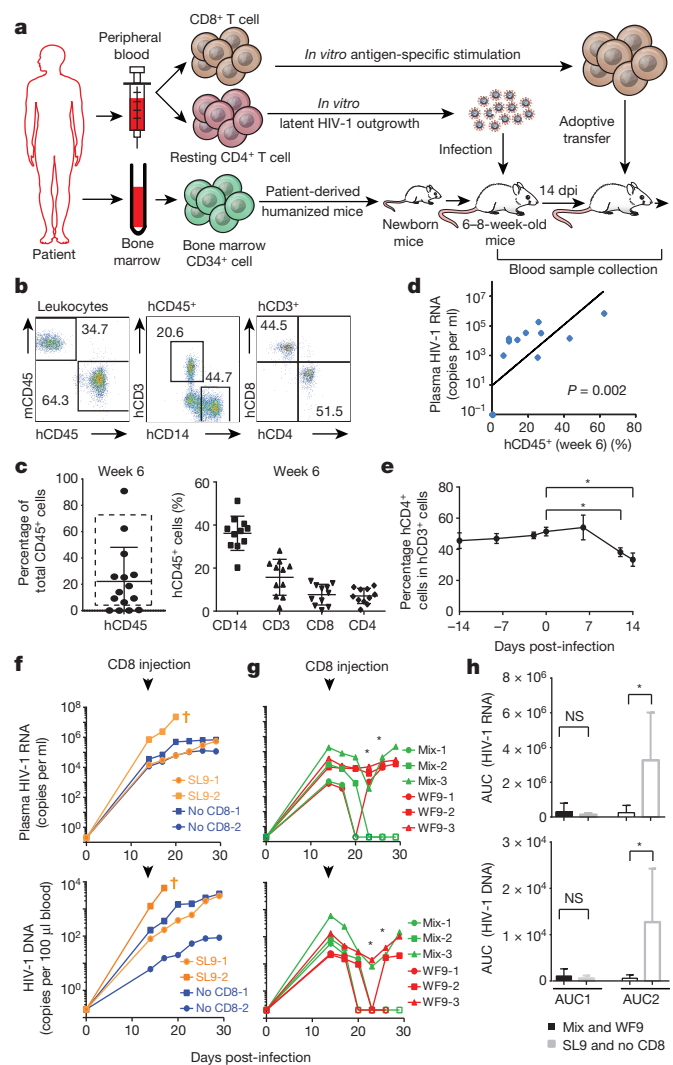
**Figure 3** | CD8<sup>+</sup> T cells targeting unmutated epitopes, not epitopes with identified escape mutations, eliminate CTL escape variants. **a**, Frequency of variants in Gag CTL epitopes in proviruses from resting CD4<sup>+</sup> T cells from patients CP36 and CP39. Epitopes tested with single-peptide stimulation are denoted in colours (red or pink, epitopes with escape observed; blue, no escape observed). **b**, Epitope-specific CD8<sup>+</sup> T cells proliferate markedly after single-peptide stimulation. Only CD8<sup>+</sup> T cells are shown. Percentages of carboxyfluorescein succinimidyl ester (CFSE)<sub>low</sub>, pentamer-positive cells are indicated for unstimulated cultures (uCD8<sup>+</sup>) with or without IL-2, for cultures stimulated with Gag peptide mixture (sCD8<sup>+</sup>) and IL-2, and for cultures stimulated with the indicated single peptides and IL-2. Wild-type versions of peptides were used for all single-peptide stimulations. **c**, CD8<sup>+</sup> T-cell proliferative responses after single-peptide stimulation. Only CD8<sup>+</sup> T cells are shown. Percentages of CFSE<sub>low</sub> cells are indicated. **d**, CD8<sup>+</sup> T cells targeting unmutated epitopes, not epitopes with identified escaped mutations, eliminate autologous CD4<sup>+</sup> T cells infected with CTL escape variants. Error bars represent s.e.m., *n* = 3. \**P* < 0.05, \*\**P* < 0.01, \*\*\**P* < 0.001, NS, not significant (*P* > 0.05), paired *t*-test.



purified CD34<sup>+</sup> cells were used to reconstitute the MIS<sup>(KI)</sup>TRG mice. We infected these patient-derived humanized mice with primary HIV-1 isolates grown from resting CD4<sup>+</sup> T cells from the same patient and then evaluated the antiviral effect of autologous CD8<sup>+</sup> T cells (Fig. 4a). MIS<sup>(KI)</sup>TRG mice engrafted with bone marrow CD34<sup>+</sup> cells from patient CP18 successfully developed human T-lymphocyte and monocyte/macrophage subsets (Fig. 4b, c), which were sufficient to support HIV-1 infection (Fig. 4d). Plasma HIV-1 RNA levels peaked 20–30 days after infection (Extended Data Fig. 6a). Depletion of CD4<sup>+</sup> T cells was clearly evident 12 days after infection in peripheral blood and spleen (Fig. 4e and Extended Data Fig. 6b, c). Cell-associated HIV-1 RNA was detected in both T cells and macrophages/monocytes (Extended Data Fig. 6d). Viral infection was also observed in various tissues in which a large number of memory CD4<sup>+</sup> T cells were detected (Extended Data Fig. 7). In control mice or mice that received autologous patient CD8<sup>+</sup> T cells pre-stimulated with a peptide representing the unmutated dominant SL9 epitope, levels of plasma HIV-1 RNA and proviral DNA in peripheral blood continued to increase from day 14 to day 29 after infection (Fig. 4f). In sharp contrast, mice that received CD8<sup>+</sup> T cells pre-stimulated with unmutated epitopes (Gag mix or WF9) had a significantly lower level of viral replication (Fig. 4g, h). Dramatic decreases in plasma HIV-1 RNA of 100- to 1,000-fold were observed in all three mice that received CD8<sup>+</sup> T cells pre-stimulated with the mixture of Gag peptides including dominant and subdominant epitopes. Two of three mice had undetectable levels of plasma HIV-1 RNA and proviral DNA in peripheral blood measured at three time points (Fig. 4g). We performed the same experiments using patient CP36-derived humanized mice and a reduction of peripheral HIV-1 RNA and DNA levels was also observed in mice that received CP36 CD8<sup>+</sup> T cells pre-stimulated with the mixture of Gag peptides (Extended Data Fig. 8). Since the post-engraftment lifespan of MIS<sup>(KI)</sup>TRG mice is only 10–12 weeks<sup>25</sup>, we were only able to investigate the acute phase of HIV-1 infection and demonstrate the *in vivo* functionality of patient CD8<sup>+</sup> T cells. Future developments of the MIS<sup>(KI)</sup>TRG model will prolong the post-engraftment lifespan of these mice and allow studies of the establishment and clearance of the HIV-1 latent reservoir *in vivo*. Together, our *in vitro* and *in vivo* experiments demonstrate that only CTL clones targeting unmutated epitopes are effective against cells infected with the viral variants that are likely to represent the major source of rebound HIV-1 after reversal of latency.

The seeding of the HIV-1 latent reservoir starts just a few days after infection<sup>26</sup>, before the development of a robust CTL response<sup>14</sup>. This is consistent with our finding that patients who initiated treatment early, in the acute infection stage, have few if any CTL escape variants archived in the latent reservoir. However, if treatment was initiated in chronic infection, CTL escape variants became dominant in the latent reservoir, indicating a complete replacement of the initially established 'wild-type' reservoir. The mechanism behind this replacement warrants further investigation, but probably reflects the dynamic nature of the reservoir in untreated infection. In any event, the overwhelming presence of escape variants in the latent reservoir of chronic patients certainly presents an additional barrier to eradication efforts. The striking difference between AP- and CP-treated patients presents another argument for early treatment of HIV-1 infection; early treatment not only reduces the size of the latent reservoir<sup>27</sup>, but also alters the composition of the reservoir, as shown here, in a way that may enhance the efficacy of potential CTL-based eradication therapies.

The hierarchy of HIV-1-specific CTL response in acute infection appears to have an important role in initial viral suppression, as demonstrated by the fact that certain immunodominant CTL populations are frequently linked to lower set point viraemia later in infection<sup>17,28</sup>. These immunodominant responses in acute infection have been identified as the major selection force driving the development of CTL escape mutations<sup>13,20</sup>. Here we show that these immunodominant response-driven mutations are not only archived in the latent reservoir, but also in fact dominate the latent provirus population in CP-treated patients.



**Figure 4 | Broad-spectrum CTLs suppress *in vivo* replication of HIV-1 from the latent reservoir of the same patients in patient-derived humanized mice.** **a**, Experimental design. dpi, days post-infection. **b**, **c**, Efficient engraftment of patient CP18-derived haematopoietic cells in MIS<sup>(KI)</sup>TRG mice at week 6. Representative flow cytometry analysis (**b**) and summary (**c**) of human CD45<sup>+</sup> cells, human T-lymphocyte and monocyte subsets. Eleven out of fifteen mice (enclosed in the rectangle) were used for HIV-1 infection. **d**, human; **m**, mouse. **e**, Correlation between frequency of peripheral human CD45<sup>+</sup> cells (6 weeks after engraftment) and plasma HIV-1 RNA levels (14 days after infection). **f**, Depletion of human CD4<sup>+</sup> T cells in peripheral blood after HIV-1 infection. \**P* < 0.05, paired *t*-test, *n* = 11. **f**, **g**, Reduction of levels of plasma HIV-1 RNA and copies of peripheral blood HIV-1 DNA after injection of viral-specific CTLs. Filled symbols, above detection limit; open symbols, below detection limit. \**P* < 0.05, unpaired *t*-test. **h**, Effect of CTLs on the level of viral replication *in vivo*. The area under the curve (AUC) of the viraemia versus time plot for each mouse in **f** (*n* = 3) and **g** (*n* = 6) before (AUC1) or after (AUC2) injection of CD8<sup>+</sup> T cells was calculated to represent quantitatively viral replication over time. Mouse SL9-2 was excluded from AUC analysis owing to early death. Error bars represent s.e.m. \**P* < 0.05, unpaired *t*-test.

Therefore, directing CTL responses to unmutated viral epitopes is essential to clear latent HIV-1. Owing to bias in antigen presentation or recognition<sup>29</sup>, common vaccination strategies will probably re-stimulate immunodominant CTL clones that do not kill infected cells after the reversal of latency. Stimulation of CTL responses with viral peptides circumvents antigen processing and is able to elicit broad-spectrum CTL responses against unmutated regions of viral proteins. Our study suggests that latent HIV-1 can be eliminated in chronically infected patients despite the overwhelming presence of CTL escape variants.



Future directions in therapeutic vaccine design need to focus on boosting broad CTL responses, as also reported elsewhere<sup>30</sup>, and/or manipulating immunodominance.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 16 December 2013; accepted 11 November 2014.**

**Published online 7 January 2015.**

1. Siliciano, J. D. *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4<sup>+</sup> T cells. *Nature Med.* **9**, 727–728 (2003).
2. Strain, M. C. *et al.* Heterogeneous clearance rates of long-lived lymphocytes infected with HIV: intrinsic stability predicts lifelong persistence. *Proc. Natl Acad. Sci. USA* **100**, 4819–4824 (2003).
3. Finzi, D. *et al.* Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
4. Chun, T. W. *et al.* Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387**, 183–188 (1997).
5. Richman, D. D. *et al.* The challenge of finding a cure for HIV infection. *Science* **323**, 1304–1307 (2009).
6. Archin, N. M. *et al.* Expression of latent HIV induced by the potent HDAC inhibitor suberoylanilide hydroxamic acid. *AIDS Res. Hum. Retroviruses* **25**, 207–212 (2009).
7. Contreras, X. *et al.* Suberoylanilide hydroxamic acid reactivates HIV from latently infected cells. *J. Biol. Chem.* **284**, 6782–6789 (2009).
8. Archin, N. M. *et al.* Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature* **487**, 482–485 (2012).
9. Shan, L. *et al.* Stimulation of HIV-1-specific cytolytic T lymphocytes facilitates elimination of latent viral reservoir after virus reactivation. *Immunity* **36**, 491–501 (2012).
10. Walker, B. D. *et al.* HIV-specific cytotoxic T lymphocytes in seropositive individuals. *Nature* **328**, 345–348 (1987).
11. Koup, R. A. *et al.* Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J. Virol.* **68**, 4650–4655 (1994).
12. Borrow, P. *et al.* Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nature Med.* **3**, 205–211 (1997).
13. Goonetilleke, N. *et al.* The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *J. Exp. Med.* **206**, 1253–1272 (2009).
14. McMichael, A. J., Borrow, P., Tomaras, G. D., Goonetilleke, N. & Haynes, B. F. The immune response during acute HIV-1 infection: clues for vaccine development. *Nature Rev. Immunol.* **10**, 11–23 (2010).
15. Phillips, R. E. *et al.* Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature* **354**, 453–459 (1991).
16. Koup, R. A. Virus escape from CTL recognition. *J. Exp. Med.* **180**, 779–782 (1994).
17. Goulder, P. J. *et al.* Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nature Med.* **3**, 212–217 (1997).
18. Goulder, P. J. & Watkins, D. I. HIV and SIV CTL escape: implications for vaccine design. *Nature Rev. Immunol.* **4**, 630–640 (2004).
19. Henn, M. R. *et al.* Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.* **8**, e1002529 (2012).
20. Liu, M. K. *et al.* Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *J. Clin. Invest.* **123**, 380–393 (2013).
21. Rolland, M., Nickle, D. C. & Mullins, J. I. HIV-1 group M conserved elements vaccine. *PLoS Pathog.* **3**, e157 (2007).
22. Létourneau, S. *et al.* Design and pre-clinical evaluation of a universal HIV-1 vaccine. *PLoS ONE* **2**, e984 (2007).
23. Yu, X. G. *et al.* Consistent patterns in the development and immunodominance of human immunodeficiency virus type 1 (HIV-1)-specific CD8<sup>+</sup> T-cell responses following acute HIV-1 infection. *J. Virol.* **76**, 8690–8701 (2002).
24. Ho, Y. C. *et al.* Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**, 540–551 (2013).
25. Rongvaux, A. *et al.* Development and function of human innate immune cells in a humanized mouse model. *Nature Biotechnol.* **32**, 364–372 (2014).
26. Whitney, J. B. *et al.* Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys. *Nature* **512**, 74–77 (2014).
27. Ananworanich, J. *et al.* Impact of multi-targeted antiretroviral treatment on gut T cell depletion and HIV reservoir seeding during acute HIV infection. *PLoS ONE* **7**, e33948 (2012).
28. Goulder, P. J. *et al.* Evolution and transmission of stable CTL escape mutations in HIV infection. *Nature* **412**, 334–338 (2001).
29. Le Gall, S., Stamegna, P. & Walker, B. D. Portable flanking sequences modulate CTL epitope processing. *J. Clin. Invest.* **117**, 3563–3575 (2007).
30. Hansen, S. G. *et al.* Cytomegalovirus vectors violate CD8<sup>+</sup> T cell epitope recognition paradigms. *Science* **340**, 1237874 (2013).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank all study participants. We thank J. Blankson for critical advice to the project; L. Alston and R. Hoh for coordinating patient recruitment; J. Alderman, C. Weibel and E. Henchey for technical assistance in the animal study. We thank the National Institutes of Health (NIH) AIDS Reagent Program for providing HIV-1 consensus B peptides. R.F.S. is supported by the Howard Hughes Medical Institute, by the Martin Delaney CARE and DARE Collaboratories (NIH grants AI096113 and 1U19AI096109), by an ARCHE Collaborative Research Grant from the Foundation for AIDS Research (amFAR 108165-50-RGRL), by the Johns Hopkins Center for AIDS Research (P30AI094189), and by NIH grant 43222. L.S. is supported by NIH grant T32 AI07019. R.A.F. is supported by the Bill and Melinda Gates Foundation and the Howard Hughes Medical Institute.

**Author Contributions** K.D., L.S., R.A.F. and R.F.S. conceived and designed the research studies; K.D., J.L., H.Z., J.B.M. and L.S. performed the *in vitro* experiments; K.D., A.R., L.W., C.G., A.J.M., D.M.V., G.D.Y., T.S., P. K. and L.S. performed animal experiments; C.M.D., G.G., H.L.M. and S.G.D. provided patient samples; K.D., M.P., L.W., H.H., J.D.S., S.L.S., L.S. and R.F.S. analysed data; K.D., L.S. and R.F.S. wrote the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.F.S. ([rsiliciano@jhmi.edu](mailto:rsiliciano@jhmi.edu)), L.S. ([liang.shan@yale.edu](mailto:liang.shan@yale.edu)) or R.A.F. ([richard.flavell@yale.edu](mailto:richard.flavell@yale.edu)).

## METHODS

**Human subjects.** Peripheral blood or bone marrow for the isolation of primary CD4<sup>+</sup>, CD8<sup>+</sup> T cells or CD34<sup>+</sup> cells was obtained from 30 HIV-1-infected patients (Extended Data Table 1) and 7 healthy adult volunteers. All patients had been on ART for at least 2 years and had maintained undetectable plasma HIV-1 RNA levels (<50 copies per ml) for at least 1 year before study. Ten AP-treated patients were recruited from the OPTIONS cohort at the University of California San Francisco (UCSF). This study was approved by the Johns Hopkins Internal Review Board and by the UCSF Committee on Human Research. Written informed consent was provided by all study participants. HLA typing for each patient was performed by the Johns Hopkins University Immunogenetics Laboratory.

**Sample preparation for deep sequencing.** Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood by Ficoll gradient separation. CD4<sup>+</sup> T cells were purified from PBMCs by negative selection using CD4<sup>+</sup> Isolation Kit II (Miltenyi). Resting CD4<sup>+</sup> T cells were then purified from CD4<sup>+</sup> T cells by negative selection using CD25, CD69 and HLA-DR microbeads (Miltenyi). Genomic DNA was extracted from 5 million resting CD4<sup>+</sup> T cells from each patient using QIAamp DNA Mini Kit (Qiagen). The *gag* gene was amplified from genomic DNA by a two-round nested PCR using these primers: 5' outer primer (5'-TTGACTAGCGGAG GCTAGAAGG-3'); 3' outer primer (5'-GATAAACCTCCAATCCCCCTAT C-3'); 5' inner primer (5'-GAGAGATGGGTGCGAGAGCGTC-3'); 3' inner primer (5'-CTGCTCTGTATCTAATAGAGC-3'). For each patient, the entire genomic DNA from 5 million resting CD4<sup>+</sup> T cells was evenly distributed as a template into 80 PCR reactions. The reactions were performed by using High Fidelity Platinum Taq Polymerase (Life Technologies) following the manufacturer's instructions. PCR amplicons were purified by gel extraction after gel electrophoresis.

**Deep sequencing.** For PacBio RS single-molecule sequencing, amplicons were bar-coded with a group of 10 bp indexes and then multiple samples were pooled together to generate a SMRTbell sequencing library following the Pacific Biosciences template preparation and sequencing-C2 user guide for 2 kb insert size and using the Pacific Biosciences DNA template preparation kit. For MiSeq sequencing, the pooled amplicon DNA was end repaired, adenylated, and ligated to Illumina TruSeq adaptors and PCR enriched for 10 cycles. The resulting library was then run on a bio-analyser high-sensitivity DNA chip for size and concentration determination. The library was then sequenced on MiSeq for paired-end 250 bp reads. The sequence reads from PacBio and MiSeq were demultiplexed using Fastx-Toolkit.

**Data analysis for deep-sequencing results.** For the paired MiSeq reads, the two reads were first merged using FLASH<sup>31</sup>. MiSeq and PacBio reads from each individual were then aligned to the reference HIV-1 consensus B Gag sequence using Bowtie2 (ref. 32). A custom program was written using Perl scripts to identify and compute the frequency of all sequence variants that caused non-synonymous amino acid changes in each individual's relevant optimal Gag epitopes (based on reported information in the HIV Molecular Immunology Database, Los Alamos National Laboratory (<http://www.hiv.lanl.gov/content/immunology/index.html>)) according to their HLA type. For each individual, variants that occurred at a frequency >3% were retained. Additionally, for PacBio reads, sequences with identified premature stop codons were eliminated from the analysed results. For each identified variation, the mutation type regarding CTL recognition was determined by matching with the information in the before-mentioned database. The five mutation types adopted in this paper are: documented escape (no CTL response when patient cells are challenged with the variant peptide); inferred escape (variant is predicted to be an escape mutant by longitudinal study or transmission study, but the reactivity of the variant is not tested experimentally); diminished response (experimental data suggest partial escape as evidenced by decreased CTL response); susceptible form (CTL response is elicited when patient cells are challenged with the variant peptide); and mutation type not determined (no experimental data on CTL recognition of this variant).

**ELISpot assays.** The ELISpot assays were performed using Human IFN- $\gamma$  ELISpot PLUS kit (Mabtech) according to methods previously described<sup>33</sup> and the manufacturer's instructions. PBMCs were added at 200,000 cells per well and synthetic peptides were added in a final concentration of 0.1, 1 or 10  $\mu\text{g ml}^{-1}$ . A response was considered positive if it was threefold higher than the mean background (cell only control) and greater than 55 SFC per million cells. The number of specific T cells was calculated by subtracting the mean background values.

**Recovery and sequencing of patient viruses from resting CD4<sup>+</sup> T cells.** Co-culture assays were performed to recover and amplify replication-competent viruses as previously described<sup>34</sup>. The viruses were recovered from 5–10 million resting CD4<sup>+</sup> T cells. The concentration of outgrowth viruses was determined by p24 ELISA (PerkinElmer). Total RNA of outgrowth viruses was extracted using TRIzol LS reagent (Life Technologies). Residual DNA was then removed by TURBO DNase (Life Technologies) treatment. First-strand complementary DNA was synthesized using SuperScript III Reverse Transcriptase (Life Technologies) and the *gag* gene was amplified from cDNA using the *gag* outer primer pair mentioned above. The

PCR amplicons were then purified by gel extraction and sequenced by regular Sanger sequencing.

**In vitro HIV-1 infection.** PBMCs from HIV-1-infected patients and healthy donors were stimulated by adding 0.5  $\mu\text{g ml}^{-1}$  PHA and IL-2 (100 U  $\text{ml}^{-1}$ ) to basal media (RPMI with 10% heat-inactivated fetal bovine serum and antibiotics) for 3 days before isolation of CD4<sup>+</sup> T cells. Each patient's activated CD4<sup>+</sup> T cells were infected with viruses recovered from the same patient's resting CD4<sup>+</sup> T cells. Healthy donors' CD4<sup>+</sup> T cells were infected with a laboratory strain virus, BaL. The virus concentration used in infection was equivalent to the p24 concentration of 200 ng  $\text{ml}^{-1}$ . All infections were performed by centrifugation of target cells with virus at 1,200g for 2 h.

**Stimulation of CD8<sup>+</sup> T cells.** PBMCs from CP-treated patients were cultured in the presence of IL-2 (100 U  $\text{ml}^{-1}$ ) with a mixture of consensus B Gag (or Nef, Rev, Tat, Env) peptides (800 ng  $\text{ml}^{-1}$  for each) (NIH AIDS Reagent Program), or with individual synthetic peptide (0.5  $\mu\text{g ml}^{-1}$ ) (Genemed Synthesis). CD8<sup>+</sup> T cells were purified after 6 days of incubation by positive selection using human CD8 microbeads (Miltenyi). To monitor CTL proliferation, PBMCs were stained with CFSE (Life Technologies) before incubation and with the relevant pentamer (Proimmune) after incubation. PBMCs were then stained with CD8-APC (Becton Dickinson (BD)) and analysed by flow cytometry using FACS Canto II (BD).

**Co-culture of autologous CD4<sup>+</sup> and CD8<sup>+</sup> T cells.** Three hours after infection, CD4<sup>+</sup> T cells were mixed with autologous unstimulated or stimulated CD8<sup>+</sup> T cells at a 1:1 ratio in basal media at 5 million cells per ml. Two days after co-culture, enfuvirtide (T-20, Roche) was added into the culture at 10  $\mu\text{M}$  to prevent further infection events except if the measurement was p24 ELISA. Three days after co-culture, cells were stained with CD8-APC (BD) first, fixed and permeabilized with Cytoperm/Cytofix (BD pharmingen), then stained for intracellular p24 Gag (PE, Coulter). Cells were analysed by flow cytometry using FACS Canto II (BD). For measurement of viral growth, 5  $\mu\text{l}$  of supernatant was taken from the co-culture at days 0, 3 and 6, and subjected to p24 ELISA. For analysis of cell contact dependence, CD4<sup>+</sup> and CD8<sup>+</sup> T cells were placed in separate chambers of *trans*-well plates (0.4  $\mu\text{m}$ , Costar).

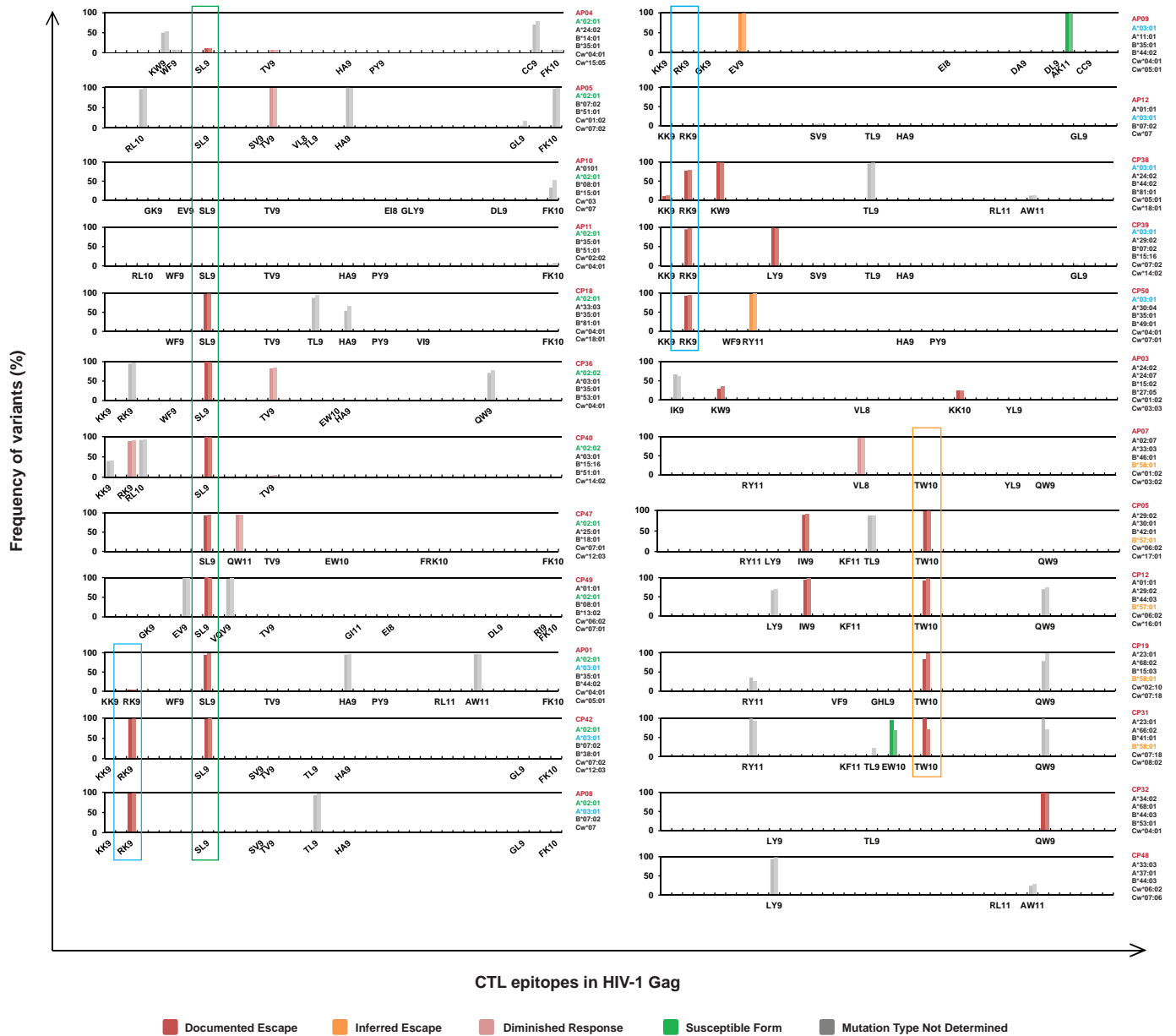
**Generation and infection of patient-derived humanized mice.** The previously reported MISTRG mouse in the *Rag2*<sup>-/-</sup> *Il2rg*<sup>-/-</sup> 129  $\times$  Balb/c (N2) genetic background harbours knock-in replacements of the endogenous mouse *Csf1*, *Csf2*, *Il3* and *Tpo* genes with humanized versions and a BAC transgene encoding human SIRP- $\alpha$ <sup>25</sup>. We generated the Sirpa<sup>(KI)</sup> mouse, which harbours a knock-in replacement of the endogenous mouse *Sirpa* gene with a humanized version. The Sirpa<sup>(KI)</sup> mouse will be thoroughly described elsewhere (manuscript in preparation). The improved MIS<sup>(KI)</sup>TRG mouse was generated by breeding Sirpa<sup>(KI)</sup> mice to MITRG mice. All animal experimentations were performed in compliance with Yale Institutional Animal Care and Use Committee protocols. MIS<sup>(KI)</sup>TRG mice were maintained with continuous treatment with enrofloxacin in the drinking water (Baytril, 0.27 mg  $\text{ml}^{-1}$ ). Patient bone marrow or fetal liver CD34<sup>+</sup> cells were isolated by CD34 microbeads selection (miltenyi). Newborn mice (within their first 3 days of life) were sublethally irradiated (X-ray irradiation with X-RAD 320 irradiator, PXI; 1  $\times$  150 cGy) and 100,000 fetal liver or 250,000 patient CD34<sup>+</sup> cells in 20  $\mu\text{l}$  of PBS were injected into the liver with a 22-gauge needle (Hamilton Company). Both male and female mice with comparable engraftment levels (percentage of hCD45<sup>+</sup>, hCD3<sup>+</sup> and hCD14<sup>+</sup> cells in the blood) were separated randomly into the experimental groups 6–8 weeks after engraftment. Mice engrafted with patient CD34<sup>+</sup> cells were infected by retro-orbital injection with HIV-1 (100 ng p24), which was recovered and expanded from the resting CD4<sup>+</sup> T cells of the same patient (CD34<sup>+</sup> cell donor), as mentioned earlier. Mice engrafted with fetal liver CD34<sup>+</sup> cells were infected by intravenous injection with HIV-1 BaL (100 ng p24). Mice with less than 5% human CD45<sup>+</sup> cells in the peripheral blood were excluded from the infection study. Mice with more than 70% human CD45<sup>+</sup> cells in the peripheral blood were also excluded because they were unhealthy due to human macrophage/monocyte-caused anaemia<sup>25</sup>. Twenty million autologous CD8<sup>+</sup> T cells with or without pre-stimulation were injected intravenously 9 or 14 days after infection. Group allocation was blinded. Peripheral blood samples were collected by retro-orbital bleeding at different time points before and after injection of CD8<sup>+</sup> T cells. Engraftment of human CD45<sup>+</sup> cells as well as lymphoid and myeloid subsets was determined by flow cytometry. Plasma HIV-1 RNA in peripheral blood was measured by one-step reverse transcriptase (Invitrogen) real-time PCR using the following primers and probe, described previously<sup>35</sup>: forward (5'  $\rightarrow$  3') ACATCAAGCAGCCATGCAAAAT, reverse (5'  $\rightarrow$  3') TCTGGCCTGGTGCAATAGG, and probe (5'  $\rightarrow$  3') VIC-CTA TCCCATTCTGCAGCTTCCTCATTGATG-TAMRA. Assay sensitivity is 200 RNA copies per ml of plasma. HIV-1 DNA in peripheral blood was also measured by real-time PCR using the same primers and probe mentioned earlier, with assay sensitivity at 5 copies per 100  $\mu\text{l}$  of blood. Total viral DNA in PBMCs was determined by measuring copies of viral DNA per 100  $\mu\text{l}$  blood and blood volume per mouse (80  $\mu\text{l}$  blood per 1 g body weight). To quantitate total viral DNA in tissues, spleens,

livers and lungs of infected mice were collected. For the spleen, single-cell suspensions were treated with ACK lysis buffer. Liver and lung leukocytes were isolated by digesting tissues with 100 U ml<sup>-1</sup> collagenase IV and 0.02 mg ml<sup>-1</sup> DNase I (Sigma), followed by density gradient centrifugation.

**Statistical analysis.** For comparison of HIV-1 variant frequency (Fig. 1b, c) and viral infection in HIV-1 BaL-infected mice (Extended Data Figs 6 and 7), we applied Mann–Whitney tests. For comparison of the inhibitory effect of autologous CTLs (Fig. 2a), we applied a Wilcoxon matched pairs test. For comparison of viral replication in humanized mice (Fig. 4f–h), we applied an unpaired *t*-test. For all other comparisons, paired *t*-tests were applied. All tests were calculated by the

GraphPad Prism 6 software, and conducted as two-tailed tests with a type I error rate of 5%. No statistical method was used to predetermine sample size.

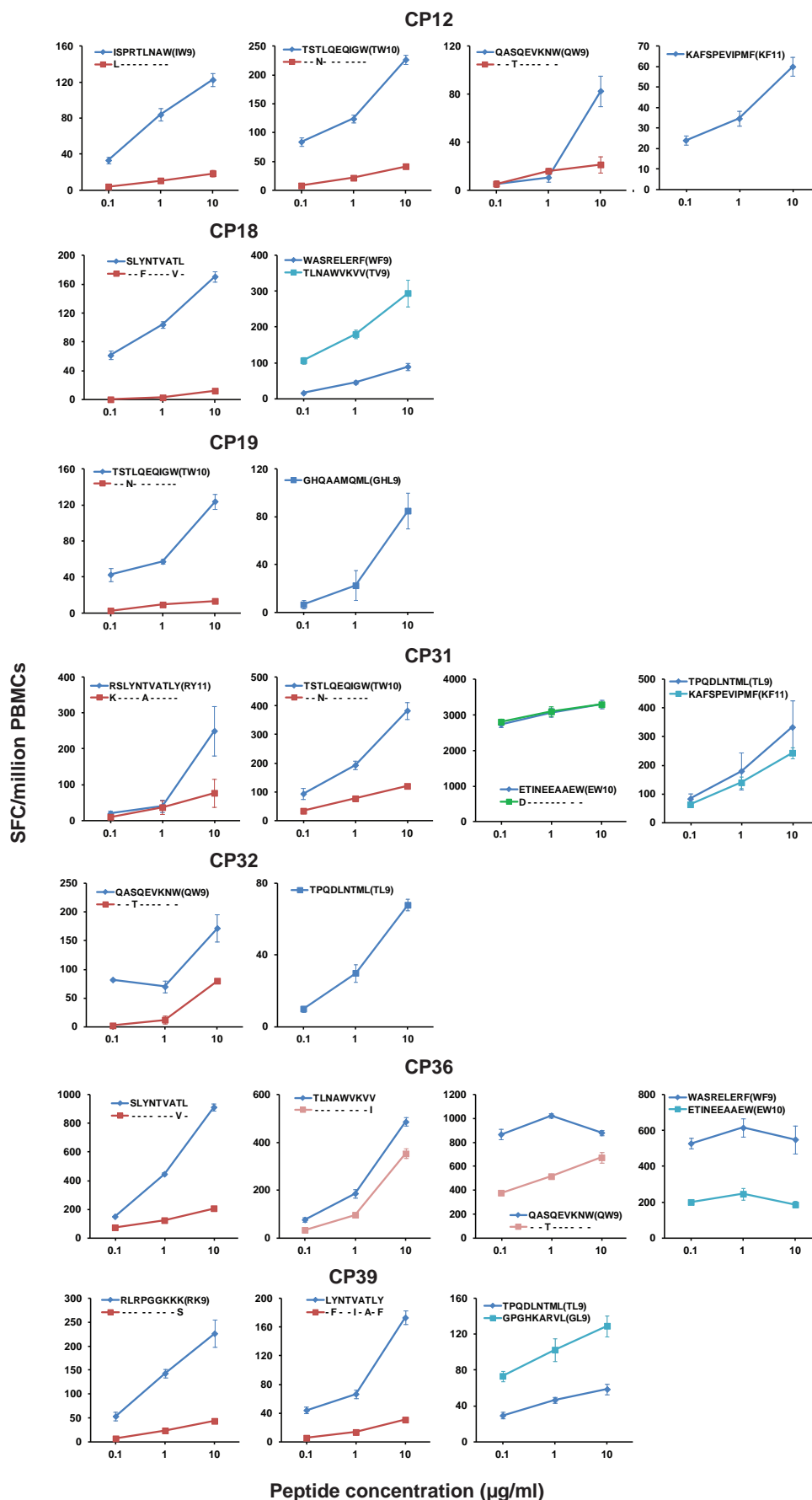
31. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
32. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
33. Streeck, H., Frahm, N. & Walker, B. D. The role of IFN- $\gamma$  Elispot assay in HIV vaccine research. *Nature Protocols* **4**, 461–469 (2009).
34. Siliciano, J. D. & Siliciano, R. F. Enhanced culture assay for detection and quantitation of latently infected, resting CD4<sup>+</sup> T-cells carrying replication-competent virus in HIV-1-infected individuals. *Methods Mol. Biol.* **304**, 3–15 (2005).



**Extended Data Figure 1 | CTL escape variants dominate the latent reservoir of CP-treated HIV-1-positive individuals, but not AP-treated individuals.** Frequency of variants in Gag CTL epitopes in proviruses from resting CD4<sup>+</sup> T cells. Results of all 25 patients tested are shown. Only optimal CTL epitopes relevant to each patient's HLA type are listed in linear positional order on the x-axis. Results from both PacBio (left bar) and MiSeq (right bar) sequencing platforms are shown for each epitope. The absence of bars above a listed epitope

indicates that only wild-type sequences were detected. For each mutation in a CTL epitope, information regarding the effect of the mutation on CTL recognition from the Los Alamos National Laboratory (LANL) HIV Molecular Immunology Database or from ELISpot assays described in Methods was used to assign the mutation to one of the categories indicated at the bottom. See Methods for definitions of these categories.



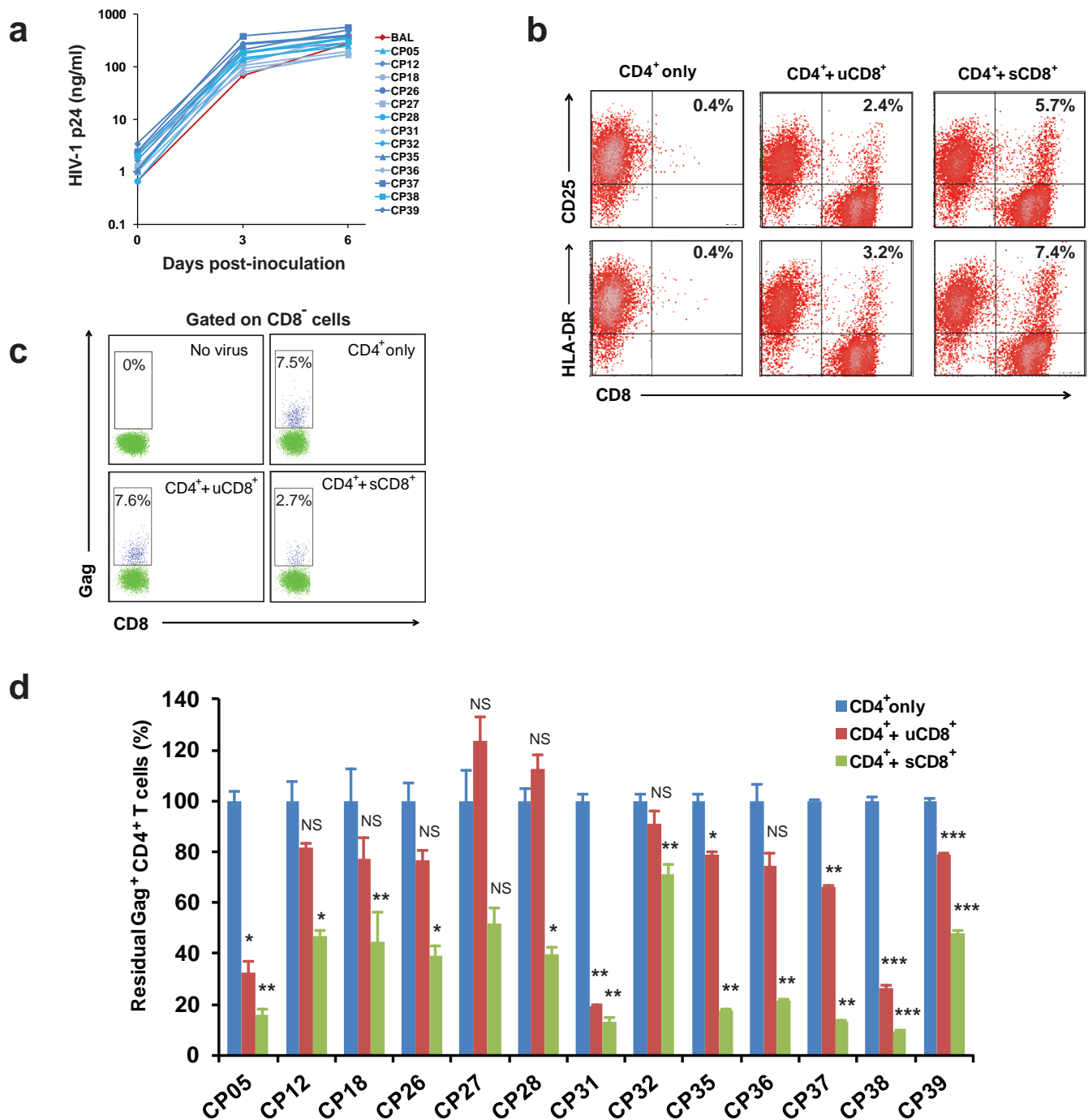


**Extended Data Figure 2 | Characterization of CTL responses against HIV-1 Gag epitopes by interferon- $\gamma$  ELISpot.** Results of seven patients tested are shown. The peptides tested are listed for each patient in each graph. Error bars represent s.e.m.,  $n = 3$ .



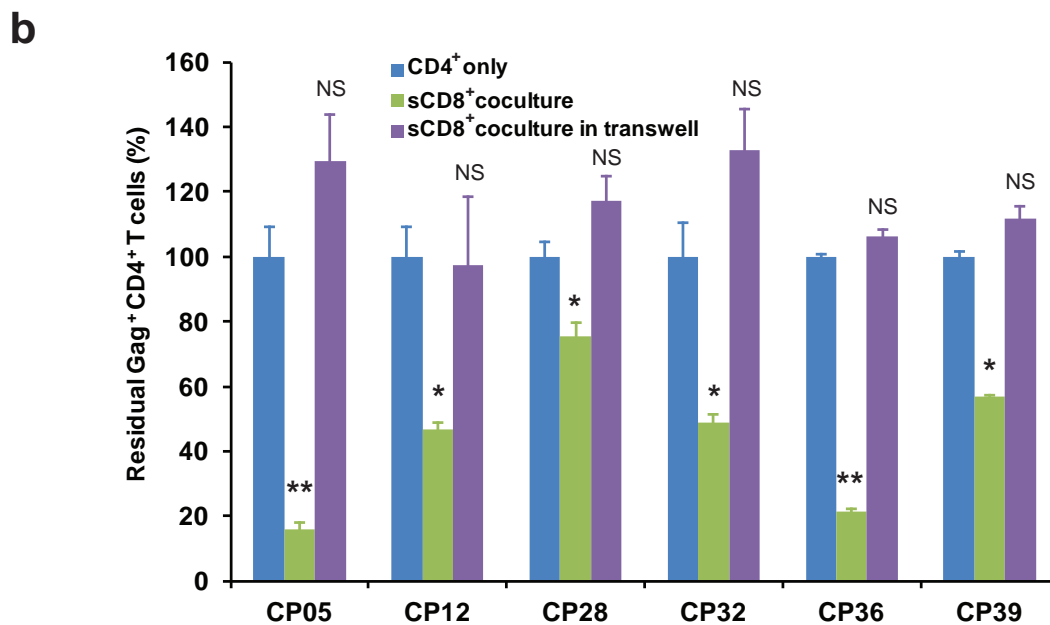
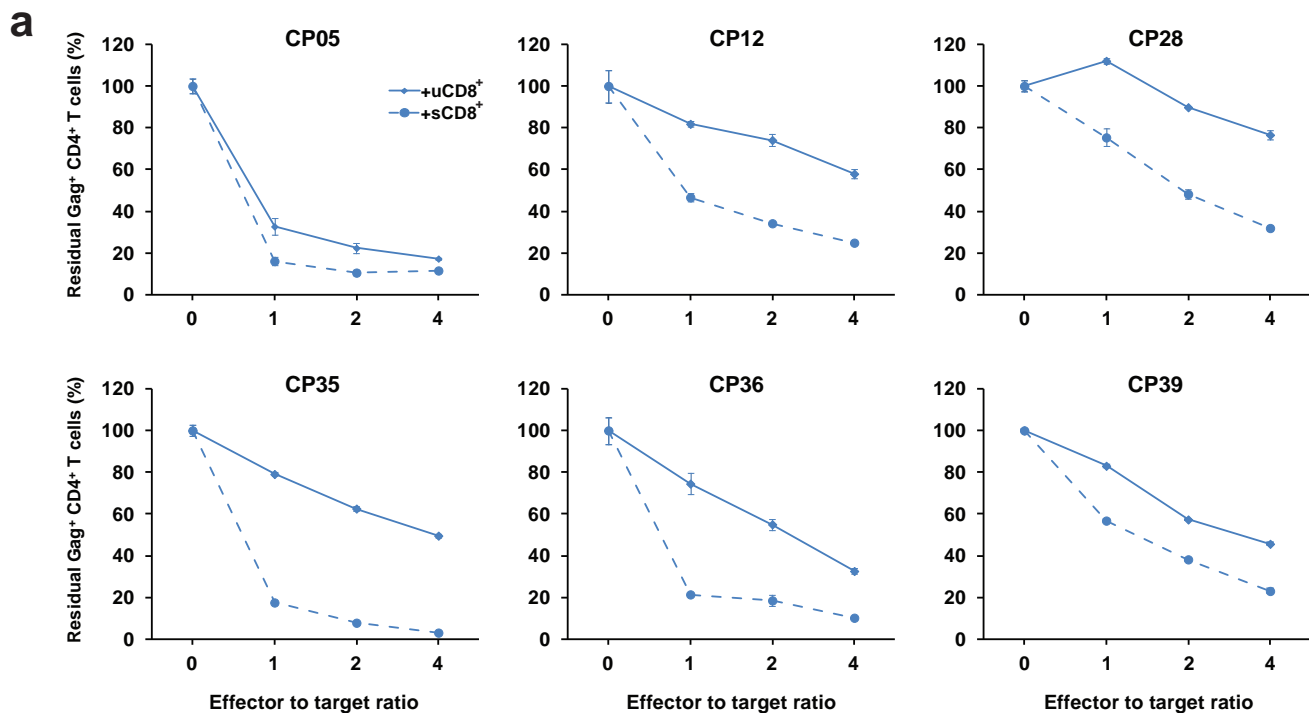
**Extended Data Figure 3 | Partial Gag sequences from proviral DNA and outgrowth virus from resting CD4<sup>+</sup> T cells from eight CP-treated patients.** CTL epitopes with no observed variation are highlighted in blue. Documented escape mutations (red shading), inferred escape mutations (yellow shading),

diminished response (pink shading), susceptible form (green shading) or undetermined variations (grey shading) in relevant optimal epitopes are indicated. See Methods for definitions of these types of mutations. This figure supplements Fig. 1e, as a total of nine CP-treated patients were tested.



**Extended Data Figure 4** | CD8<sup>+</sup> T cells pre-stimulated with a mixture of consensus B Gag peptides eliminate autologous CD4<sup>+</sup> T cells infected with autologous HIV-1 from resting CD4<sup>+</sup> T cells. **a**, HIV-1 isolated from ART-treated individuals replicates as well as the laboratory strain virus Bal. p24 values represent mean of three replicates. Error bars represent s.e.m.,  $n = 3$ . **b**, CD8<sup>+</sup> T cells are not stimulated after co-culture with PHA-activated CD4<sup>+</sup> T cells. **c**, A representative flow cytometric analysis of CTL-mediated killing

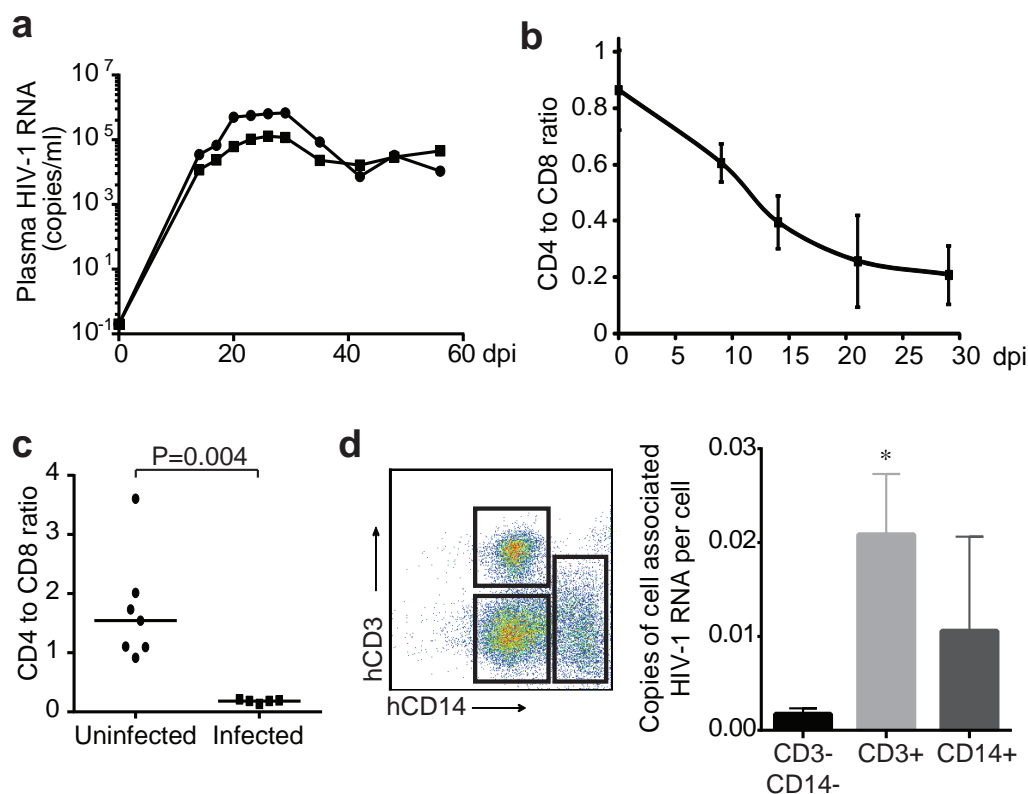
after co-culture of infected CD4<sup>+</sup> T cells with autologous CD8<sup>+</sup> T cells. CTL activity is measured by the percentage of Gag-positive, CD8-negative cells after 3 days of co-culture relative to cultures without CD8<sup>+</sup> T cells. **d**, Pre-stimulated CD8<sup>+</sup> T cells eliminate autologous infected CD4<sup>+</sup> T cells more efficiently than non-stimulated CD8<sup>+</sup> T cells. All results were normalized to the CD4 only control group. Error bars represent s.e.m.,  $n = 3$ . \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , NS, not significant ( $P > 0.05$ ), paired  $t$ -test.



**Extended Data Figure 5 | The elimination of infected CD4<sup>+</sup> T cells is mediated by direct killing by autologous CD8<sup>+</sup> T cells. a,** Killing of infected CD4<sup>+</sup> T cells is enhanced by increased effector to target ratios for both pre-stimulated and non-stimulated CD8<sup>+</sup> T cells. **b,** Killing of the infected

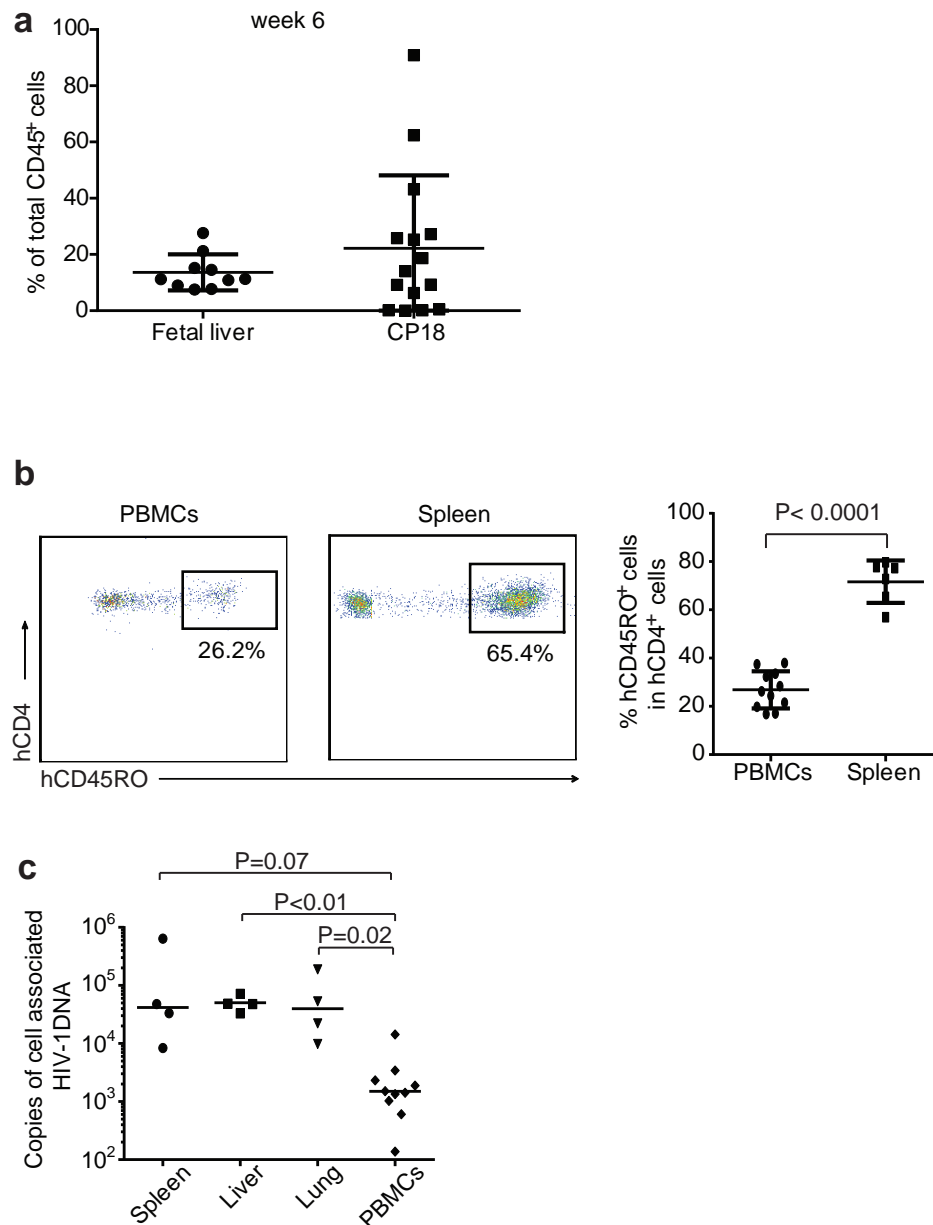
CD4<sup>+</sup> T cells depends on direct cell–cell contact between CD4<sup>+</sup> T cells and CTLs. All results were normalized to the CD4<sup>+</sup> only control group. Error bars represent s.e.m.,  $n = 3$ . \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , NS, not significant ( $P > 0.05$ ), paired  $t$ -test.





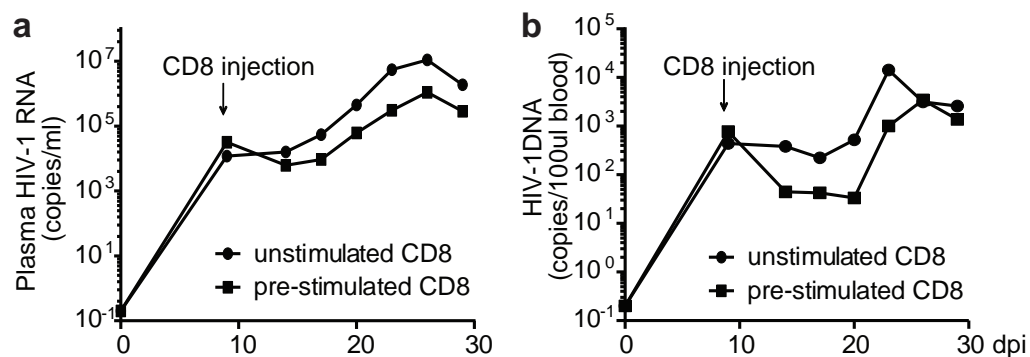
**Extended Data Figure 6 | Viral dynamics and depletion of CD4<sup>+</sup> T cells in humanized mice.** **a**, Viral dynamics in CP18-infected MIS<sup>(KI)</sup>TRG mice. CP18-derived MIS<sup>(KI)</sup>TRG mice were infected with autologous HIV-1. Plasma HIV-1 RNA levels were measured from day 0 to day 56. **b**, Depletion of CD4<sup>+</sup> T cells in peripheral blood of HIV-1 BaL-infected mice. MIS<sup>(KI)</sup>TRG mice engrafted with fetal liver CD34<sup>+</sup> cells were infected with HIV-1 BaL. The CD4 to CD8 ratio in peripheral blood was measured by fluorescence-activated cell sorting (FACS) from day 0 to day 29 after infection. Error bars represent s.e.m.,  $n = 5$ . **c**, Depletion of CD4<sup>+</sup> T cells in spleen of HIV-1 BaL-infected

mice. MIS<sup>(KI)</sup>TRG mice engrafted with fetal liver CD34<sup>+</sup> cells were infected with HIV-1 BaL. The CD4 to CD8 ratio in spleen was measured by FACS 20 days after infection. Medians and  $P$  values from Mann–Whitney test are shown. **d**, Detection of cell-associated HIV-1 RNA in T cells and macrophages/monocytes. CD3<sup>+</sup> and CD14<sup>+</sup> human cells from HIV-1-infected MIS<sup>(KI)</sup>TRG mice from spleen and lung were purified by FACS. CD3<sup>−</sup>CD14<sup>−</sup> cells were also collected as controls. Cell-associated HIV-1 RNA was quantified by *gag*-specific quantitative polymerase chain reaction (qPCR). Error bars represent s.e.m.,  $n = 3$ . \* $P < 0.05$ , unpaired  $t$ -test.



**Extended Data Figure 7 | HIV-1 infection occurs in peripheral blood and tissues in humanized mice.** **a**, Engraftment levels of MIS<sup>(KI)</sup>TRG mice with fetal liver or patient CD34<sup>+</sup> cells. **b**, Memory CD4<sup>+</sup> T cells are detected in MIS<sup>(KI)</sup>TRG mice after infection. MIS<sup>(KI)</sup>TRG mice were infected with HIV-1 BaL. Peripheral blood and indicated tissues from infected mice were

collected at 20 days post-infection. Memory CD4<sup>+</sup> T cells were determined by CD45RO staining. **c**, Total number of cell-associated HIV-1 DNA in blood and tissues. DNA from peripheral blood or indicated tissues was isolated for the measurement of total amount of cell-associated HIV-1 DNA by real-time PCR. **b**, **c**, Medians and *P* values from Mann–Whitney test are shown.



**Extended Data Figure 8 | Broad-spectrum CTLs suppress *in vivo* infection of patient-derived humanized mice with autologous latent HIV-1.** The generation of patient CP36-derived humanized mice is described in Fig. 4. Mice were infected with autologous viruses at 6 weeks old. CD8<sup>+</sup> T cells from patient

CP36 were pre-stimulated with the mixture of Gag peptides or left untreated for 6 days *in vitro*, and were injected into mice intravenously 9 days after infection. Plasma HIV-1 RNA (**a**) and HIV-1 DNA (**b**) in peripheral blood were measured by real-time PCR.

Extended Data Table 1 | Characteristics of study subjects

Patient	Year of Diagnosis	CD4 count* (cells/ $\mu$ l)	Plasma HIV-1 RNA <sup>†</sup> (copies/ml)	Time on ART (years)	Treatment start time after infection (days)
AP01	2006	1251	<50	7	64
AP03	2004	595	<50	9	34
AP04	1998	953	<50	15	77
AP05	2002	618	<50	11	39
AP07	2012	592	<50	1.5	67
AP08	2008	780	<50	6	28
AP09	2012	1069	<50	2	39
AP10	2006	513	<50	7	50
AP11	2007	874	<50	6	10
AP12	2007	629	<50	6	15
CP05	2001	500	<50	10	>180
CP12	1997	1074	<50	15	>180
CP18	1998	773	<50	>4	>180
CP19	2006	620	<50	6	>180
CP26	1994	640	<50	16	>180
CP27	1987	784	<50	4	>180
CP28	1998	614	<50	6	>180
CP31	2000	619	<50	10	>180
CP32	1999	780	<50	2	>180
CP35	2002	738	<50	10	>180
CP36	2003	1119	<50	4	>180
CP37	1999	730	<50	12	>180
CP38	1986	870	<50	3	>180
CP39	1996	1152	<50	10	>180
CP40	2002	544	<50	7	>180
CP42	1987	684	<50	16	>180
CP47	1986	792	<50	14	>180
CP48	1998	641	<50	8	>180
CP49	1992	864	<50	5	>180
CP50	2001	964	<50	4	>180

\*Patient CD4 count was measured during this study.

<sup>†</sup>Plasma HIV-1 RNA levels for all patients were <50 copies per ml for at least 1 year before this study.



# CEACAM1 regulates TIM-3-mediated tolerance and exhaustion

Yu-Hwa Huang<sup>1</sup>, Chen Zhu<sup>2\*</sup>, Yasuyuki Kondo<sup>1\*</sup>, Ana C. Anderson<sup>2</sup>, Amit Gandhi<sup>1</sup>, Andrew Russell<sup>3</sup>, Stephanie K. Dougan<sup>4</sup>, Britt-Sabina Petersen<sup>5</sup>, Espen Melum<sup>1,6</sup>, Thomas Pertel<sup>2</sup>, Kiera L. Clayton<sup>7</sup>, Monika Raab<sup>8</sup>, Qiang Chen<sup>9</sup>, Nicole Beauchemin<sup>10</sup>, Paul J. Yazaki<sup>11</sup>, Michal Pyzik<sup>1</sup>, Mario A. Ostrowski<sup>7,12</sup>, Jonathan N. Glickman<sup>13</sup>, Christopher E. Rudd<sup>8</sup>, Hidde L. Ploegh<sup>4</sup>, Andre Franke<sup>5</sup>, Gregory A. Petsko<sup>3</sup>, Vijay K. Kuchroo<sup>2</sup> & Richard S. Blumberg<sup>1</sup>

T-cell immunoglobulin domain and mucin domain-3 (TIM-3, also known as HAVCR2) is an activation-induced inhibitory molecule involved in tolerance and shown to induce T-cell exhaustion in chronic viral infection and cancers<sup>1–5</sup>. Under some conditions, TIM-3 expression has also been shown to be stimulatory. Considering that TIM-3, like cytotoxic T lymphocyte antigen 4 (CTLA-4) and programmed death 1 (PD-1), is being targeted for cancer immunotherapy, it is important to identify the circumstances under which TIM-3 can inhibit and activate T-cell responses. Here we show that TIM-3 is co-expressed and forms a heterodimer with carcinoembryonic antigen cell adhesion molecule 1 (CEACAM1), another well-known molecule expressed on activated T cells and involved in T-cell inhibition<sup>6–10</sup>. Biochemical, biophysical and X-ray crystallography studies show that the membrane-distal immunoglobulin-variable (IgV)-like amino-terminal domain of each is crucial to these interactions. The presence of CEACAM1 endows TIM-3 with inhibitory function. CEACAM1 facilitates the maturation and cell surface expression of TIM-3 by forming a heterodimeric interaction in *cis* through the highly related membrane-distal N-terminal domains of each molecule. CEACAM1 and TIM-3 also bind in *trans* through their N-terminal domains. Both *cis* and *trans* interactions between CEACAM1 and TIM-3 determine the tolerance-inducing function of TIM-3. In a mouse adoptive transfer colitis model, CEACAM1-deficient T cells are hyper-inflammatory with reduced cell surface expression of TIM-3 and regulatory cytokines, and this is restored by T-cell-specific CEACAM1 expression. During chronic viral infection and in a tumour environment, CEACAM1 and TIM-3 mark exhausted T cells. Co-blockade of CEACAM1 and TIM-3 leads to enhancement of anti-tumour immune responses with improved elimination of tumours in mouse colorectal cancer models. Thus, CEACAM1 serves as a heterophilic ligand for TIM-3 that is required for its ability to mediate T-cell inhibition, and this interaction has a crucial role in regulating autoimmunity and anti-tumour immunity.

We examined the role of CEACAM1 in ovalbumin (OVA)-specific peripheral T-cell tolerance<sup>11</sup>. OVA protein administration (Extended Data Fig. 1a) resulted in tolerance induction in wild-type OVA-specific T-cell receptor transgenic OT-II *Rag2*<sup>−/−</sup> mice (Fig. 1a), but not in transgenic OT-II *Ceacam1*<sup>−/−</sup> *Rag2*<sup>−/−</sup> mice (Fig. 1b). Transfer of carboxy-fluorescein diacetate succinimidyl ester (CFSE)-labelled naive CD4<sup>+</sup> Vα2<sup>+</sup> T cells from transgenic OT-II *Rag2*<sup>−/−</sup> mice into *Ceacam1*<sup>−/−</sup> recipients (Extended Data Fig. 1b, c) was associated with increased OVA induced proliferation (Fig. 1c) and TIM-3 expression uniformly restricted to proliferating CEACAM1<sup>+</sup> T cells (Fig. 1d and Extended Data Fig. 1d).

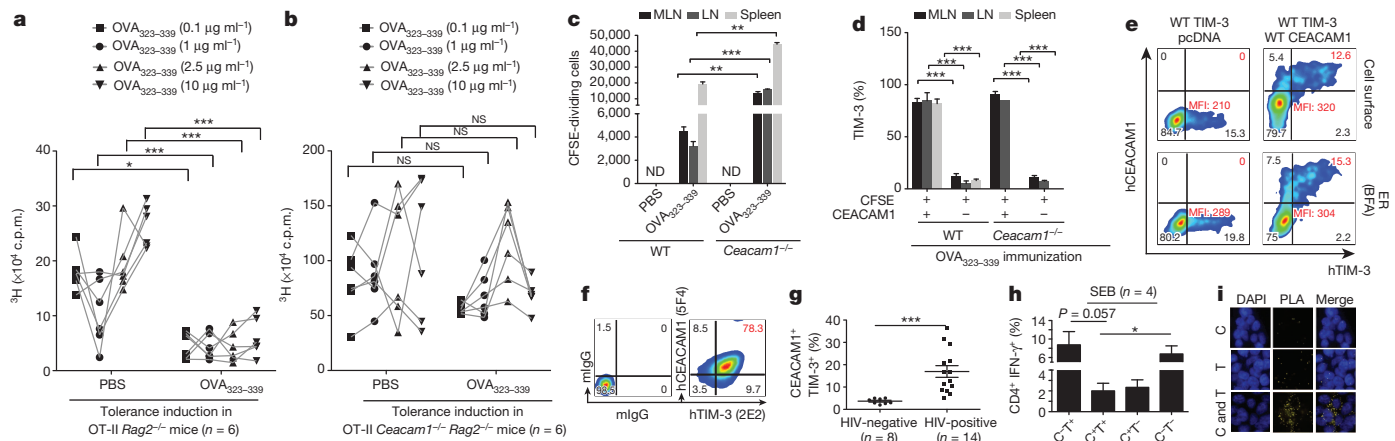
CD4<sup>+</sup> T-cell receptor (TCR) Vβ8<sup>+</sup> T cells in *Ceacam1*<sup>−/−</sup>, but not wild-type, mice lacked TIM-3 expression after *Staphylococcus aureus* enterotoxin B (SEB) administration, suggesting CEACAM1 and TIM-3 co-expression on tolerized T cells (Extended Data Fig. 1e, f). Flag-tagged human (h) CEACAM1 enhanced cell surface expression of co-transfected haemagglutinin (HA)-tagged hTIM-3 in human embryonic kidney 293T (HEK293T) cells, with virtually all hTIM-3-positive HEK293T cells notably CEACAM1-positive (Fig. 1e). Human T cells co-expressed TIM-3 and CEACAM1 after *in vitro* activation with decreased CEACAM1 expression after *TIM3* (also known as *HAVCR2*) silencing (Fig. 1f and Extended Data Fig. 1g, h). Human immunodeficiency virus (HIV)-infected, but not uninfected, subjects exhibited increased CEACAM1<sup>+</sup> TIM-3<sup>+</sup> (double-positive) CD4<sup>+</sup> T cells, which were poor producers of interferon-γ (IFN-γ), as were double-positive CD8<sup>+</sup> T cells (Fig. 1g, h and Extended Data Fig. 1i–l). *In situ* proximity ligation analysis<sup>12</sup> of hCEACAM1 and hTIM-3 co-transfected HEK293T cells (Fig. 1i and Extended Data Fig. 1m–o), and co-cultures of activated primary human T cells (Extended Data Fig. 1p, q) confirmed the nearness of both molecules on the cell surface of HEK293T cells and co-localization within the immune synapse of activated T cells, respectively.

TIM-3 has been proposed to engage an unknown ligand<sup>13</sup> (Extended Data Fig. 2a–c), and we considered CEACAM1 a possible candidate that is known to homodimerize<sup>14</sup>. Modelling available X-ray crystallographic structures of mouse (m) CEACAM1 (ref. 14) and mTIM-3 (ref. 13) membrane-distal IgV-like, N-terminal domains predicted structural similarity with extensive interactions along their FG–CC' interface in *cis* and *trans* configurations (Extended Data Fig. 2d–g and Supplementary Information). Mouse T-cell lymphoma cells predicted to possess a novel TIM-3 ligand expressed CEACAM1 (refs 13, 15) (Extended Data Fig. 2h, i). hCEACAM1, but not integrin α5 (ITGA5) (Extended Data Fig. 3a), was co-immunoprecipitated with hTIM-3 and vice-versa from co-transfected HEK293T cells (Fig. 2a, b). Co-immunoprecipitation of CEACAM1 and TIM-3 was confirmed with activated primary human T cells (Extended Data Fig. 3b) and primary mouse T cells from either *Ceacam1*-4L<sup>Tg</sup> *Ceacam1*<sup>−/−</sup> mice<sup>6</sup> or *Ceacam1*-4S<sup>Tg</sup> *Ceacam1*<sup>−/−</sup> mice<sup>10</sup> (transgenic mice in which CEACAM1 isoforms containing a long (L) or short (S) cytoplasmic tail, respectively, are conditionally overexpressed in T cells)<sup>7</sup> (Extended Data Fig. 3c).

Although tunicamycin treatment had no effect (Extended Data Fig. 3d), mutation of amino acid residues including natural human allelic variants anticipated to be involved in these interactions disrupted the co-immunoprecipitation of hCEACAM1 and hTIM-3 in co-transfected

<sup>1</sup>Division of Gastroenterology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, Massachusetts 02115, USA. <sup>2</sup>Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Harvard Institutes of Medicine, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. <sup>3</sup>Rosenstiel Basic Medical Sciences Research Center, Brandeis University, 415 South Street, Waltham, Massachusetts 02454, USA. <sup>4</sup>Whitehead Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. <sup>5</sup>Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel 24105, Germany. <sup>6</sup>Norwegian PSC Research Center, Division of Cancer Medicine, Surgery and Transplantation, Oslo University Hospital, Oslo 0424, Norway. <sup>7</sup>Department of Immunology, University of Toronto, Toronto, Ontario M5S1A8, Canada. <sup>8</sup>Cell Signalling Section, Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK. <sup>9</sup>State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu 610041, China. <sup>10</sup>Goodman Cancer Research Centre, McGill University, Montreal H3G 1Y6, Canada. <sup>11</sup>Beckman Institute, City of Hope, Duarte, California 91010, USA. <sup>12</sup>Keenan Research Centre of St. Michael's Hospital, Toronto, Ontario M5S1A8, Canada. <sup>13</sup>GI Pathology, Miraca Life Sciences, Newton, Massachusetts 02464, USA.

\*These authors contributed equally to this work.



**Figure 1 | TIM-3 and CEACAM1 are co-expressed on T cells during induction of tolerance.** **a, b**, Tolerance induction in indicated mice. Median c.p.m., counts per minute. **c, d**, Responses of CFSE-labelled transgenic OT-II *Rag2*<sup>-/-</sup> T cells in mesenteric lymph nodes (MLN), peripheral lymph node (LN) or spleen of wild-type (WT) or *Ceacam1*<sup>-/-</sup> recipients to PBS (*n* = 3 per group) or OVA (*n* = 5 per group) for proliferation (**c**) and CEACAM1 or TIM-3 (**d**) expression. ND, not detectable. **e**, hCEACAM1 and hTIM-3 expression in co-transfected HEK293T cells. Percentage and mean fluorescence intensity (MFI) of hTIM-3 indicated. BFA, brefeldin A; ER, endoplasmic

reticulum. **f**, hCEACAM1 and hTIM-3 expression on activated primary CD4<sup>+</sup> human T cells. **g**, hCEACAM1<sup>+</sup> TIM-3<sup>+</sup> CD4<sup>+</sup> T cells (**g**) and intracellular cytokine staining for IFN-γ in CD4<sup>+</sup> T cells after SEB stimulation (**h**) in HIV infection. C, CEACAM1; T, TIM-3 (*n* = 4 per group). **i**, *In situ* proximity ligation assay of hCEACAM1 and hTIM-3 co-transfected HEK293T as in **e**. DAPI, 4',6-diamidino-2-phenylindole. All data are mean ± s.e.m. and represent five (**e, f**), three (**c, d, i**) and two (**a, b**) independent experiments. \**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.001.

HEK293T cells (Fig. 2c, d, Extended Data Fig. 3e–s, Extended Data Table 1 and Supplementary Information).

A single-chain of the hCEACAM1 N-terminal domain (amino acids 1–107) joined to the hTIM-3 N-terminal domain (amino acids 1–105) by a linker (GGGGS)<sub>4</sub> with a hexahistidine tag appended to the carboxy terminus was expressed in *Escherichia coli* (Extended Data Fig. 4a), and shown to interact specifically with the N-terminal domain of hTIM-3 by surface plasmon resonance (Extended Data Fig. 4b–e). This single-chain protein was crystallized and a structural model built from X-ray diffraction data (Extended Data Table 2). This revealed two similar copies of a hCEACAM1 (IgV domain)–hTIM-3 (IgV domain) heterodimer interacting along each of the respective FG–CC' faces (Fig. 2e) and at amino acid contact points demonstrated to be involved in biochemical interactions between hCEACAM1 and hTIM-3 (Fig. 2f, g and Extended Data Fig. 4f–h), with no conformational changes of the hCEACAM1 IgV domain upon forming a heterodimer with hTIM-3 in comparison to a CEACAM1 homodimer (Extended Data Table 2, Extended Data Fig. 4i–k and Supplementary Information).

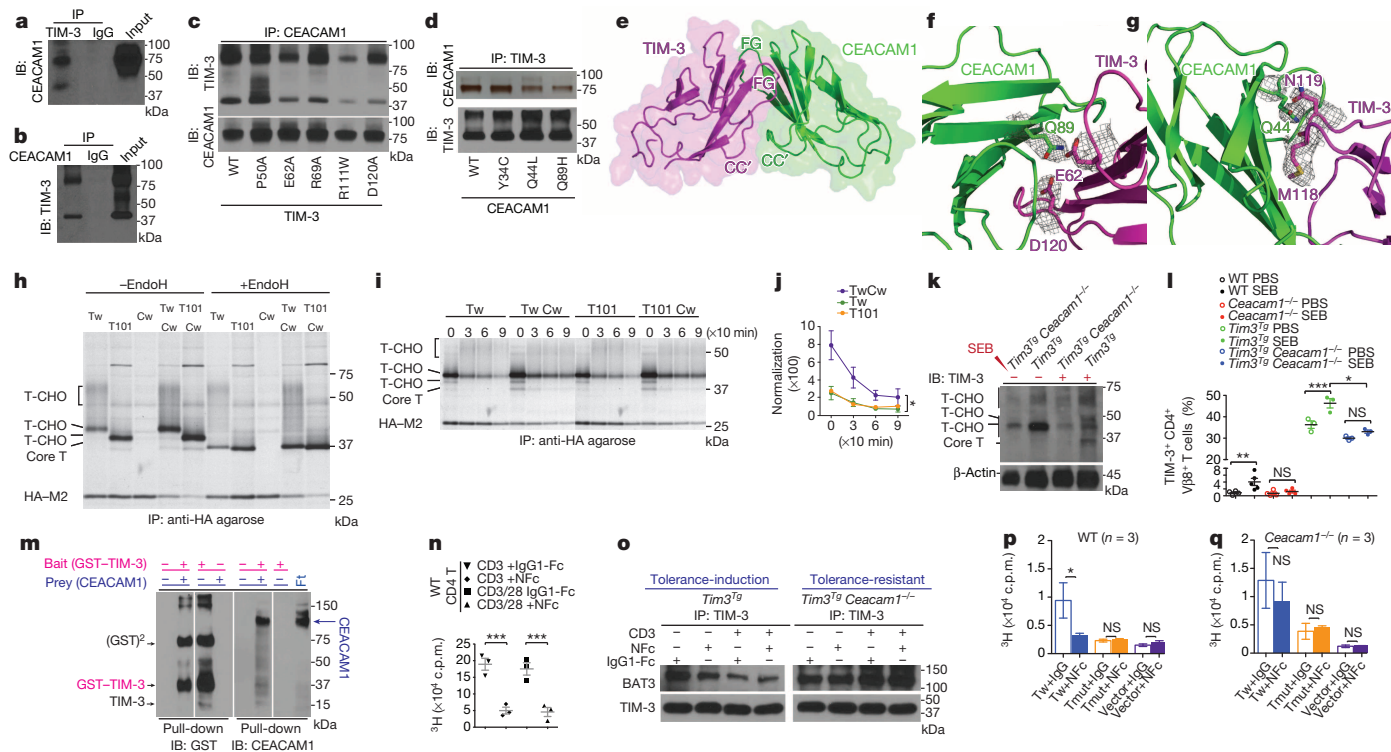
Metabolic labelling (Fig. 2h–j) showed overall enhancement of hTIM-3 biosynthesis after hCEACAM1 co-expression with decreased cell surface display, intracellular accumulation and hypoglycosylation of hTIM-3 when mutant, hypomorphic forms of hTIM-3 or hCEACAM1 were co-expressed in HEK293T cells (Extended Data Fig. 5a–e). Especially profound was a natural hTIM-3 variant (rs147827860 (Thr101Ile)) that although exhibiting normal core protein association with hCEACAM1 (Extended Data Fig. 5f) could not be rescued by hCEACAM1 co-expression and remained hypoglycosylated with increased intracellular retention and a near absence of cell surface expression (Fig. 2h, i, Extended Data Fig. 5a, b, e and Supplementary Information).

CD4<sup>+</sup> T cells from SEB-treated *Ceacam1*<sup>-/-</sup> mice or mice with transgenic overexpression of *Tim3* (also known as *Havcr2*) in T cells<sup>16</sup> lacking CEACAM1 expression (*Tim3*<sup>Tg</sup> *Ceacam1*<sup>-/-</sup>) exhibited diminished expression of all mTIM-3 forms (fully glycosylated, hypoglycosylated and core protein) (Fig. 2k), decreased upregulation of TIM-3 expression on the cell surface (Fig. 2l) and blunted deletion of CD4<sup>+</sup> Vβ8<sup>+</sup> T cells (Extended Data Fig. 5g) relative to wild-type or transgenic *Tim3*<sup>Tg</sup> mice expressing endogenous CEACAM1. T-helper 1 (T<sub>H</sub>1) polarized cells from *Tim3*<sup>Tg</sup> *Ceacam1*<sup>-/-</sup> mice were resistant to galectin-9-induced apoptosis (Extended Data Fig. 5h), which is dependent on TIM-3 expression as a glycoprotein<sup>15</sup>.

The hTIM-3 N-terminal domain bound hCEACAM1 and stained the cell surface of human CEACAM1-4L-transfected Jurkat T cells (Fig. 2m and Extended Data Fig. 5i, j). *Trans*-ligation of wild-type mouse CD4<sup>+</sup> T cells with a mCEACAM1 N-terminal domain–Fc fusion protein (NFC) rendered them unresponsive to *in vitro* stimulation with anti-CD3 and anti-CD28 (Fig. 2n and Extended Data Fig. 5k). CD4<sup>+</sup> T cells from SEB-treated *Tim3*<sup>Tg</sup> or *Tim3* *Ceacam1*<sup>-/-</sup> mice exhibited similar levels of TIM-3 associated BAT3, a repressor of TIM-3 cytoplasmic tail signalling<sup>17</sup>, at the time of T-cell isolation (Fig. 2o). However, only T cells from SEB-tolerized *Tim3*<sup>Tg</sup> mice, but not SEB-treated *Tim3*<sup>Tg</sup> *Ceacam1*<sup>-/-</sup> mice that were resistant to SEB-induced tolerance, exhibited BAT3 dissociation from mTIM-3 after NFC fusion protein ligation with further release observed after addition of anti-CD3 (Fig. 2o and Extended Data Fig. 5l, m). We observed similar levels of NFC fusion protein staining of primary T cells from *Tim3*<sup>Tg</sup> and *Tim3*<sup>Tg</sup> *Ceacam1*<sup>-/-</sup> mice, relative to that observed with *Ceacam1*<sup>-/-</sup> mice, which was blocked by a TIM-3-specific antibody (Extended Data Fig. 5n), implicating heterophilic ligation of TIM-3 as the major factor responsible for T-cell inhibition and BAT3 release.

Highly activated CD4<sup>+</sup> T cells from wild-type and *Ceacam1*<sup>-/-</sup> mice were transduced with a retrovirus encoding green fluorescent protein (GFP) plus mTIM-3. Anti-CD3 induced proliferation and TNF-α secretion was inhibited by NFC fusion protein ligation in transduced wild-type but not *Ceacam1*<sup>-/-</sup> T cells or all groups transduced with GFP alone (Fig. 2p, q and Extended Data Fig. 5o, p). Both wild-type and *Ceacam1*<sup>-/-</sup> T-cell transductants with a mTIM-3 cytoplasmic tail deletion construct (TIM-3<sup>Δ252–281</sup>)<sup>17,18</sup> were unresponsive to anti-CD3 stimulation or NFC fusion protein ligation (Fig. 2p, q and Extended Data Fig. 5o, p).

Tolerance pathways are important in T-cell regulation<sup>19</sup>. We detected homozygous rs147827860 (Thr101Ile) carriage in inflammatory bowel disease (IBD) but not controls (Extended Data Table 3a–c). We thus transferred naive CD4<sup>+</sup> CD62L<sup>high</sup> CD44<sup>+</sup> T cells from wild-type mice into *Ceacam1*<sup>-/-</sup> *Rag2*<sup>-/-</sup> recipients and observed colon-infiltrating CEACAM1<sup>+</sup> TIM-3<sup>+</sup> T cells expressing markedly decreased intracellular levels of IFN-γ, IL-2 and IL-17A relative to other lamina propria T cells (Fig. 3a, b). Colon-infiltrating lamina propria T cells from adoptive transfer of naive *Ceacam1*<sup>-/-</sup> CD4<sup>+</sup> T cells into *Ceacam1*<sup>-/-</sup> *Rag2*<sup>-/-</sup> recipients expressed reduced cell surface levels of TIM-3 (Fig. 3c) and increased intracellular levels of TNF-α (Fig. 3d) in association with a severe progressive colitis with increased mortality (Fig. 3e, dagger symbol), and histopathological evidence of injury (Fig. 3f) characterized by



**Figure 2 | CEACAM1 and TIM-3 heterodimerize and serve as heterophilic ligands.** **a, b,** Co-immunoprecipitation (IP) and immunoblot (IB) of wild-type hCEACAM1 and hTIM-3 in co-transfected HEK293T cells. **c, d,** Co-immunoprecipitation and immunoblot of wild-type hCEACAM1 and hTIM-3 mutants (**c**) or wild-type hTIM-3 and hCEACAM1 mutants (**d**) as in **a** and **b**. **e,** Human CEACAM1 (IgV)-TIM-3 (IgV) heterodimer structure. **f, g,**  $2F_o - F_c$  maps contoured at  $0.9\sigma$  showing electron densities. **h, i,** Autoradiogram of anti-haemagglutinin (HA) (hTIM-3) immunoprecipitate from metabolically-labelled (**h**) and pulse-chase metabolically-labelled (**i**) co-transfected HEK293T cells. CHO, carbohydrate; core T, non-glycosylated hTIM-3; Cw, wild-type hCEACAM1; EndoH, endoglycosidaseH; H2-MA, HA-tagged influenza virus A M2 protein; T, hTIM-3(Thr101Ile); Tw, wild-type hTIM-3. hTIM-3 isoforms noted. **j,** Quantification of densities in **i** ( $n = 3$  per group). **k,** Immunoblot for mTIM-3

acute and chronic inflammation (Extended Data Fig. 6a). Reestablishment of T-cell-specific mCEACAM1-4L expression using *Ceacam1-4L<sup>Tg</sup>* *Ceacam1<sup>-/-</sup>* mice<sup>6</sup>, restored TIM-3 display and decreased intracellular TNF- $\alpha$  expression by the infiltrating lamina propria T cells together with decreased disease severity (Fig. 3c–f and Extended Data Fig. 6b), reflective of CEACAM1<sup>+</sup> TIM-3<sup>+</sup> T-cell restoration.

CD4<sup>+</sup> T cells from *Tim3<sup>Tg</sup> Ceacam1<sup>-/-</sup>* mice, in contrast to those from CEACAM1-proficient *Tim3<sup>Tg</sup>* mice, were hyper-inflammatory after transfer into *Ceacam1<sup>-/-</sup> Rag2<sup>-/-</sup>* recipients and caused significant weight loss and severe colitis with increased expression of T<sub>H</sub>17 signature genes (Fig. 3g–i and Extended Data Fig. 6c, d). Nanostring (Extended Data Fig. 6e) and quantitative PCR (Fig. 3j) analysis of lamina propria mononuclear cells demonstrated increased transcripts for *Ebi3*, *IL-27p28* (also known as *Il27*) and *IL-12p35* (also known as *Il12a*) encoding the regulatory cytokines IL-27 and IL-35 (refs 20, 21) in *Ceacam1<sup>-/-</sup> Rag2<sup>-/-</sup>* recipients of CD4<sup>+</sup> T cells from *Tim3<sup>Tg</sup>* relative to wild-type mice, and their absence in recipients of *Ceacam1<sup>-/-</sup>* or *Tim3<sup>Tg</sup> Ceacam1<sup>-/-</sup>* T cells.

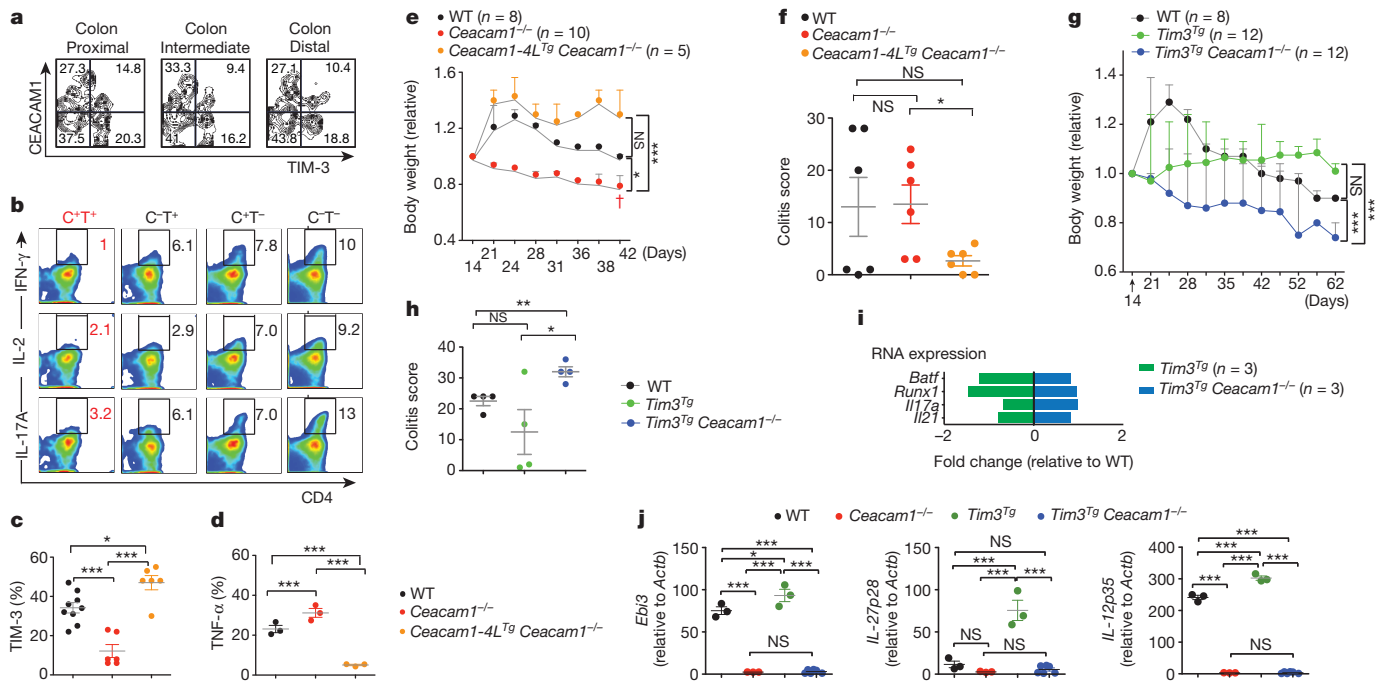
Anti-tumour immunity is hindered by T-cell expression of inhibitory molecules associated with an exhausted phenotype<sup>22</sup>. We observed high levels of CEACAM1<sup>+</sup> TIM-3<sup>+</sup>, relative to PD-1<sup>+</sup> TIM-3<sup>+</sup>, CD4<sup>+</sup> and CD8<sup>+</sup> T cells at the tumour site of wild-type mice in the azoxymethane (AOM)/dextran sodium sulphate (DSS) model of colitis-associated colon cancer (Extended Data Fig. 6f–j). At 2.5% DSS, *Tim3<sup>Tg</sup>* mice exhibited 100% fatality, significantly greater than all other genotypes (Fig. 4a). Tumour quantification at a lower DSS dose (1.5%) showed that the polyp

from PBS-treated (–) or SEB-treated (+) CD4<sup>+</sup> T cells. Labelling as in **h** and **i**. **l,** mTIM-3 expression after SEB tolerance induction. **m,** Column-bound glutathione S-transferase (GST)-hTIM-3 IgV-domain pull-down of hCEACAM1 detected by immunoblot. GST<sup>2</sup>, GST-hTIM-3 dimer. Ft, flow through. **n,** Suppression of mouse CD4<sup>+</sup> T-cell proliferation by mCEACAM1 N-terminal domain-Fc fusion protein (NfC). **o,** Immunoprecipitation of mTIM-3 and immunoblot for BAT3 or mTIM-3 from lysates of CD4<sup>+</sup> T cells. **p, q,** Proliferation of CD4<sup>+</sup> T cells from wild-type (**p**) and *Ceacam1<sup>-/-</sup>* (**q**) mice transduced with wild-type mTIM-3 (Tw), mTIM-3 $\Delta$ 252–281 (Tmut) or vector exposed to anti-CD3 and either NfC or IgG1-Fc (IgG1). Data are mean  $\pm$  s.e.m. and represent five (**a, b**), four (**c, d**), three (**h–j, l, n, p, q**) and two (**k, m, o**) independent experiments. NS, not significant; \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ .

numbers, polyp size, high-grade dysplasia and frank carcinoma were significantly higher in *Tim3<sup>Tg</sup>* than in wild-type mice with decreased neoplasia observed in *Tim3<sup>Tg</sup> Ceacam1<sup>-/-</sup>* and *Ceacam1<sup>-/-</sup>* mice (Fig. 4b–d and Extended Data Fig. 6k).

After implantation of colorectal cancer cells (CT26) subcutaneously in BALB/c mice, a significant fraction of CD8<sup>+</sup> tumour infiltrating lymphocytes (TILs) was triple-positive T cells (TIM-3<sup>+</sup> PD-1<sup>+</sup> CEACAM1<sup>+</sup>), with the relative cell surface levels of TIM-3 correlating with CEACAM1 (Fig. 4e). PD-1<sup>+</sup> TIM-3<sup>bright</sup> CEACAM1<sup>+</sup> T cells were characterized by extremely low intracellular IL-2 and TNF- $\alpha$  expression consistent with exhaustion (Fig. 4f). Co-administration of an anti-mCEACAM1 and anti-mTIM-3 antibody delayed subcutaneous tumour growth in a preventative model (Fig. 4g and Extended Data Fig. 7a), and was able to exceed protection with co-blockade of PD-1 and TIM-3 (Extended Data Fig. 7b, c). In a therapeutic model, blockade of CEACAM1 synergized with PD-1 inhibition (Extended Data Fig. 7d, e). CEACAM1 and TIM-3 co-blockade was associated with increased CD8<sup>+</sup> (Fig. 4h) and CD4<sup>+</sup> (Fig. 4i) TILs with enhanced IFN- $\gamma$  production (Fig. 4j) and decreased IL-10 production (Extended Data Fig. 7f) by the infiltrating CD8<sup>+</sup> and CD4<sup>+</sup> T cells, respectively. Increased tumour antigen-specific CD8<sup>+</sup> T cells defined by AH1-tetramer staining<sup>23</sup>, relative to the total CD8<sup>+</sup> T-cell population (Fig. 4k and Extended Data Fig. 7g) was observed in the draining lymph nodes that correlated with tumour growth inhibition (Extended Data Fig. 7h). CT26 tumour growth was impeded in *Ceacam1<sup>-/-</sup>* mice (Fig. 4l) in association with increased AH-1 tetramer<sup>+</sup>





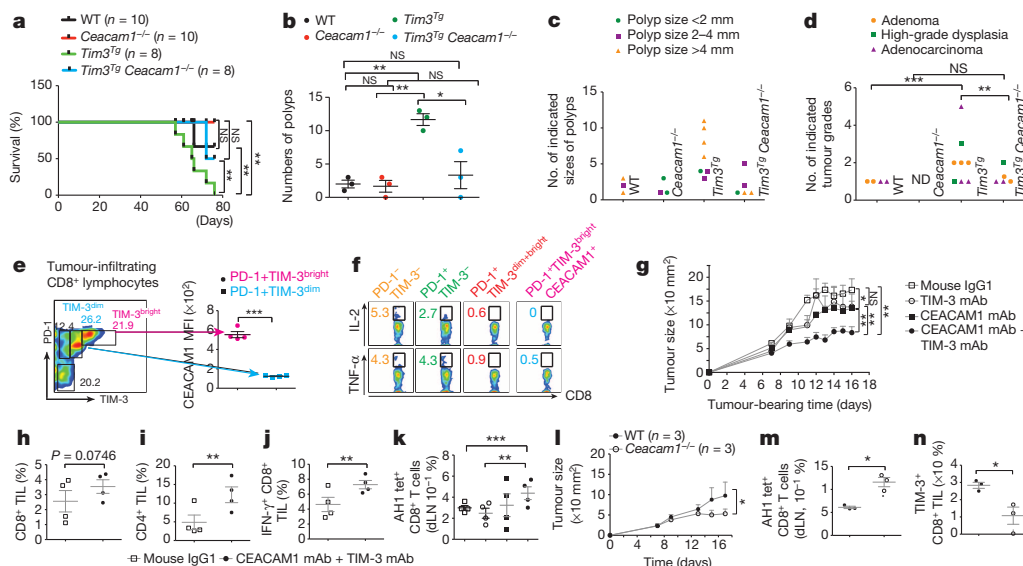
**Figure 3 | TIM-3 regulation of mucosa-associated inflammation requires CEACAM1.** **a**, mCEACAM1 and mTIM-3 expression on colonic lamina propria CD4<sup>+</sup> T cells. **b**, Intracellular cytokines in cells described in **a**. **c**, **d**, mTIM-3 (**c**) and intracellular TNF- $\alpha$  (**d**) expression in lamina propria CD4<sup>+</sup> T cells from indicated donors. **e**, Body weights relative to weights on day 14 of groups in **c** and **d**. Five mice expired ( $\dagger$ ). **f**, Score of surviving mice

of groups in **e**, **g**. Body weights of genotypes as in **e**, **h**, Score of groups described in **g**, **i**, **j**, Nanostring (**i**) and quantitative PCR (**j**) of lamina propria mononuclear cells. *Actb*,  $\beta$ -actin gene. All data are mean  $\pm$  s.e.m. and represent six (**a**), four (**b**) and three (**c**–**j**) independent experiments. \* $P$  < 0.05; \*\* $P$  < 0.01; \*\*\* $P$  < 0.001.

CD8<sup>+</sup> T cells in draining lymph nodes (Fig. 4m and Extended Data Fig. 7i) and decreased TIM-3 expression on tumour-associated TILs (Fig. 4n).

In conclusion, we show that CEACAM1 and TIM-3 form a new heterodimeric complex that is determined by interactions between their structurally similar membrane-distal IgV-like, N-terminal domains. This

association is evident during early biosynthesis and crucial for proper maturation, cell surface display and function of TIM-3 and probably CEACAM1 when expressed together. CEACAM1 also functions as an essential *trans*-heterophilic ligand for TIM-3 with the tolerance-inducing functions of TIM-3 requiring interactions with CEACAM1 in both *cis* and *trans* configurations. CEACAM1 expression with TIM-3 is further



**Figure 4 | CEACAM1 determines TIM-3 regulation of anti-tumour immune responses.** **a**, Survival curves in AOM/2.5% DSS model. **b**–**d**, Assessment of polyp numbers (**b**), polyp size (**c**) and cancer grades (**d**) in AOM/1.5% DSS model. **e**, Staining of CD8<sup>+</sup> T cells associated with CT26 tumours. **f**, Intracellular cytokine expression in TIL subsets after anti-CD3 stimulation. **g**, Prevention of CT26 tumour growth in wild-type mice ( $n$  = 5 per group). mAb, monoclonal antibody. **h**–**k**, Analysis of TILs for relative

proportion of CD8<sup>+</sup> (**h**) and CD4<sup>+</sup> (**i**) T cells, IFN- $\gamma$ <sup>+</sup> CD8<sup>+</sup> T cells (**j**) and tumour-specific (AH1-tetramer, tet<sup>+</sup>) CD8<sup>+</sup> T cells in draining lymph nodes (dLN) (**k**) in groups described in **g**. **l**–**n**, Growth of CT26 cells (**l**), AH1 tet<sup>+</sup> CD8<sup>+</sup> T cells in dLN (**m**) and TIM-3 expression on TILs (**n**) in wild-type and *Ceacam1*<sup>-/-</sup> mice. Data are mean  $\pm$  s.e.m. and represent four (**e**), three (**g**–**k**) and two (**a**–**d**, **f**, **l**–**n**) independent experiments. \* $P$  < 0.05; \*\* $P$  < 0.01; \*\*\* $P$  < 0.001.



characteristic of T cells with an exhausted phenotype such that in the absence of these interactions inflammatory responses are unrestrained and anti-tumour immunity enhanced. Together, these studies describe a novel class of heterodimeric protein interactions that function in tolerance induction with broad implications for many types of infectious, autoimmune and neoplastic conditions.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 December 2013; accepted 8 September 2014.

Published online 26 October 2014.

- Monney, L. *et al.* Th1-specific cell surface protein Tim-3 regulates macrophage activation and severity of an autoimmune disease. *Nature* **415**, 536–541 (2002).
- Sabatos, C. A. *et al.* Interaction of Tim-3 and Tim-3 ligand regulates T helper type 1 responses and induction of peripheral tolerance. *Nature Immunol.* **4**, 1102–1110 (2003).
- Sánchez-Fueyo, A. *et al.* Tim-3 inhibits T helper type 1-mediated auto- and alloimmune responses and promotes immunological tolerance. *Nature Immunol.* **4**, 1093–1101 (2003).
- Jones, R. B. *et al.* Tim-3 expression defines a novel population of dysfunctional T cells with highly elevated frequencies in progressive HIV-1 infection. *J. Exp. Med.* **205**, 2763–2779 (2008).
- Sakuishi, K. *et al.* Targeting Tim-3 and PD-1 pathways to reverse T cell exhaustion and restore anti-tumor immunity. *J. Exp. Med.* **207**, 2187–2194 (2010).
- Nagaishi, T. *et al.* SHP1 phosphatase-dependent T cell inhibition by CEACAM1 adhesion molecule isoforms. *Immunity* **25**, 769–781 (2006).
- Gray-Owen, S. D. & Blumberg, R. S. CEACAM1: contact-dependent control of immunity. *Nature Rev. Immunol.* **6**, 433–446 (2006).
- Iijima, H. Specific regulation of T helper cell 1-mediated murine colitis by CEACAM1. *J. Exp. Med.* **199**, 471–482 (2004).
- Boulton, I. C. & Gray-Owen, S. D. Neisserial binding to CEACAM1 arrests the activation and proliferation of CD4<sup>+</sup> T lymphocytes. *Nature Immunol.* **3**, 229–236 (2002).
- Chen, L. *et al.* The short isoform of the CEACAM1 receptor in intestinal t cells regulates mucosal immunity and homeostasis via Tfh cell induction. *Immunity* **37**, 930–946 (2012).
- Kearney, E. R., Pape, K. A., Joh, D. Y. & Jenkins, M. K. Visualization of peptide-specific T cell immunity and peripheral tolerance induction *in vivo*. *Immunity* **1**, 327–339 (1994).
- Söderberg, O. *et al.* Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nature Methods* **3**, 995–1000 (2006).
- Cao, E. *et al.* T cell immunoglobulin mucin-3 crystal structure reveals a galectin-9-independent ligand-binding surface. *Immunity* **26**, 311–321 (2007).
- Tan, K. *et al.* Crystal structure of murine sCEACAM1a[1,4]: a coronavirus receptor in the CEA family. *EMBO J.* **21**, 2076–2086 (2002).
- Zhu, C. *et al.* The Tim-3 ligand galectin-9 negatively regulates T helper type 1 immunity. *Nature Immunol.* **6**, 1245–1252 (2005).
- Dardalhon, V. *et al.* Tim-3/galectin-9 pathway: regulation of Th1 immunity through promotion of CD11b<sup>+</sup>Ly-6G<sup>+</sup> myeloid cells. *J. Immunol.* **185**, 1383–1392 (2010).
- Rangachari, M. *et al.* Bat3 promotes T cell responses and autoimmunity by repressing Tim-3-mediated cell death and exhaustion. *Nature Med.* **18**, 1394–1400 (2012).

- Lee, J. *et al.* Phosphotyrosine-dependent coupling of Tim-3 to T-cell receptor signaling pathways. *Mol. Cell. Biol.* **31**, 3963–3974 (2011).
- Barber, D. L. *et al.* Restoring function in exhausted CD8 T cells during chronic viral infection. *Nature Cell Biol.* **439**, 682–687 (2006).
- Hirahara, K. *et al.* Interleukin-27 priming of T cells controls IL-17 production in *trans* via induction of the ligand PD-L1. *Immunity* **36**, 1017–1030 (2012).
- Collison, L. W. *et al.* The inhibitory cytokine IL-35 contributes to regulatory T-cell function. *Nature* **450**, 566–569 (2007).
- Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nature Rev. Cancer* **12**, 298–306 (2012).
- Huang, A. Y. *et al.* The immunodominant major histocompatibility complex class I-restricted antigen of a murine colon tumor derives from an endogenous retroviral gene product. *Proc. Natl Acad. Sci. USA* **93**, 9730–9735 (1996).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank T. Gallagher, M. Yoshida and K. Holmes for essential reagents, R. Gali for statistical assistance, C. Chen, T. Wesse, S. Sabet, S. Greve, T. Henke, D. Tan, K. Sakuishi and J. Sullivan for technical assistance, E. Greenfield and C. Bencsics for core services, and J. H. Wang, E. Reinherz, R. Grenha, H. Iijima, J. Shively, A. Kaser, T. E. Adolph, K. Baker, D. Ringe and S. Zeissig for discussions. We thank the staff of the Dana Farber/Harvard Cancer Center monoclonal antibody core for purification of proteins used in X-ray crystallography and beam line X25 and X6A of the National Synchrotron Light Source (NSLS), Brookhaven National Laboratory, USA. The NSLS is supported by the US Department of Energy. This work was supported by the American Cancer Society grant RSG-11-057-01-LIB (A.C.A.); the Norwegian PSC research center and the Unger Vetlesen Medical Fund (E.M.); Crohn's & Colitis Foundation of America fellowship grant (Y.-H.H.); Deutsche Forschungsgemeinschaft (DFG) Cluster of Excellence 'Inflammation at Interfaces' Award (A.F. and B.-S.P.); Harvard Clinical Translational Science Center, UL1 TR001102 (R. Gali); the National Basic Research Program of China No. 2010CB529906 (Q.C.); Canadian Institute of Health Research (K.L.C. and N.B.); Canadian Institute of Health Research grant MOP-93787 (M.A.O.); AACR-Pancreatic Cancer Action Network (H.L.P. and S.K.D.); National Institutes of Health (NIH) grant GM32415 (G.A.P.); NIH grants AI073748, NS045937, AI039671 and AI056299 (V.K.K.); NIH grants DK044319, DK051362, DK053056, DK088199, the Harvard Digestive Diseases Center (HDDC) DK0034854 and High Point Foundation (R.S.B.).

**Author Contributions** Y.-H.H., C.Z. and Y.K. performed most experiments and helped prepare the manuscript. B.-S.P., E.M. and A.F. provided expertise in the genetic assessment for TIM-3. J.N.G. assessed all pathology. A.C.A. designed and directed tumour experiments. T.P. designed shRNA experiments. M.R. and C.E.R. performed proximity ligation analysis. S.K.D. and H.L.P. conducted and analysed pulse-chase biosynthetic labelling experiments. A.G., A.R., Q.C. and G.A.P. performed X-ray crystallography or structural analysis. K.L.C. and M.A.O. conducted immune synapse experiments. M.P. and P.J.Y. assisted with the single chain protein analysis. N.B. assisted in generation of *Ceacam1*<sup>-/-</sup> *Rag2*<sup>-/-</sup> mice and in analyses of data. R.S.B. and V.K.K. devised and coordinated the project, and together with Y.-H.H., C.Z. and Y.K. wrote the manuscript and designed the experiments. R.S.B. and V.K.K. share senior authorship on this paper.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.S.B. ([rblumberg@partners.org](mailto:rblumberg@partners.org)).

## METHODS

**Mice.** C57BL/6 and BALB/c mice were purchased from Jackson Laboratory. *Ceacam1*<sup>-/-</sup> mice (C57BL/6 and BALB/c backcrosses) were generated by N. Beauchemin<sup>24</sup>. Transgenic OT-II *Ceacam1*<sup>-/-</sup> *Rag2*<sup>-/-</sup> mice were generated by first crossing *Ceacam1*<sup>-/-</sup> mice to *Rag2*<sup>-/-</sup> (Taconic) for >6 generations to obtain *Ceacam1*<sup>-/-</sup> *Rag2*<sup>-/-</sup> double-knockout mice on a C57BL/6 background, followed by crossing these mice to transgenic OT-II *Rag2*<sup>-/-</sup> mice (Taconic) for >9 generations. *Tim3*<sup>Tg</sup> (ref. 16) and *Ceacam1*<sup>-/-</sup> mice on a C57BL/6 background were intercrossed to generate *Tim3*<sup>Tg</sup> *Ceacam1*<sup>-/-</sup> mice. *Ceacam1*-4L<sup>Tg</sup> *Ceacam1*<sup>-/-</sup> and *Ceacam1*-4S<sup>Tg</sup> *Ceacam1*<sup>-/-</sup> mice have been previously described<sup>6,10</sup>. All studies were performed in mice with a C57BL/6 genetic background except for CT26 tumour experiments that were in a BALB/c genetic background. Animal studies were conducted in a gender and age-matched manner for all experiments. Both male and female mice were used and were 6–8 weeks of age at the time of experiments or 6–10 weeks for the adoptive transfer colitis experiments. The number of animals used per group was based on previous experimental results and observed variability. Histopathology analysis was performed in a blinded manner by an expert pathologist (J.N.G.). For all other *in vitro* and *in vivo* analyses, investigators were not blinded to treatment allocation. All experiments were approved and conducted according to the guidelines set forth by the Harvard Medical Area Standing Committee on Animals.

**Genotyping.** Mouse tails were digested in tail lysis buffer (100 mM Tris-HCl, pH 8.5, 5 mM EDTA, 0.2% SDS, 200 mM NaCl and proteinase K (Roche)) overnight at 55 °C. Genomic DNA was phenol-extracted and isopropanol precipitated. DNA was dissolved in TE buffer. The primer sequences for genotyping are available on request.

**T-cell tolerance animal models.** For SEB tolerance induction model, animals of the indicated genotypes were injected intraperitoneally with 25 µg of SEB (Millipore) on days 0 and 4, and euthanized at day 8 after the first injection. Peripheral lymph node cells were isolated and stained for T-cell receptor Vβ6, Vβ8 and CD4, and analysed for TIM-3 and CEACAM1 expression on CD4<sup>+</sup> Vβ8<sup>+</sup> cells. Cells were also re-stimulated *in vitro* at 0.5 × 10<sup>6</sup> cells per ml with anti-CD3 (0.1–10.0 µg ml<sup>-1</sup>) and assessed for proliferation by [<sup>3</sup>H]-thymidine (1 µCi ml<sup>-1</sup>) uptake or production of IL-2 by ELISA. For OVA antigen-specific T-cell tolerance induction, high-dose OVA antigen was used and modified according to previous studies<sup>11,25,26</sup>. In brief, transgenic OT-II *Rag2*<sup>-/-</sup> or *Ceacam1*<sup>-/-</sup> *Rag2*<sup>-/-</sup> mice were injected intraperitoneally with 500 µg of OVA peptide (323–339, 323-ISQAVHAHAHAINEAGR-329, AnaSpec) in an equal volume amount of complete Freund's adjuvant (CFA) or PBS alone at day 0 (activation of immune response) and at day 4 (tolerance induction). At day 10, CD4<sup>+</sup> T cells were isolated from spleens and pooled peripheral lymphocytes from each experimentally treated mouse as indicated above. The pooled lymphocytes from each mouse were re-stimulated *in vitro* with various concentrations of OVA<sub>323–339</sub> peptide (0.1, 1, 2.5 and 10 µg ml<sup>-1</sup>). The 96-well cultures were set up with a total cell number of 2.5 × 10<sup>4</sup> in 200 µl RPMI complete medium pulsed with 1 µCi [<sup>3</sup>H]-thymidine (PerkinElmer Life Sciences) for the final 8 h of the 72 h assay, and collected with a Packard Micromate cell harvester. Counts per minute (c.p.m.) were determined using a Packard Matrix 96 direct counter (Packard Biosciences) and data expressed as median values. For *in vivo* tracking of OVA-specific antigen-specific T-cell responses, adoptive transfer experiments were performed on the basis of previous studies<sup>27</sup>. In brief, CD4<sup>+</sup> Vα2<sup>+</sup> T cells were isolated from spleens and pooled peripheral lymphocytes of transgenic OT-II *Ceacam1*<sup>+/+</sup> (wild-type) *Rag2*<sup>-/-</sup> mice and the cells were pre-incubated for 10 min with 4 µM CFSE in PBS plus 0.1% FBS. CFSE-labelled cells (5 × 10<sup>5</sup>) were injected intravenously by tail vein into syngeneic C57BL/6 wild-type or *Ceacam1*<sup>-/-</sup> recipients. After 18 h to allow the adoptively transferred T cells to establish themselves *in vivo*, and which showed no differences in parking of the T cells among the various genotypes, the mice were then immunized with OVA<sub>323–339</sub> (50 µg) in an equal volume of CFA. On day 3, cells were collected from spleen, mesenteric lymph and peripheral lymph node of individual mice and stained with DAPI to exclude dead cells, and concurrently stained with anti-CD4-PerCp, anti-CEACAM1- allophycocyanin and anti-TIM-3-phycoerythrin. Total lymphocytes were gated and the CFSE subset representing the transgenic OT-II, adoptively transferred CD4<sup>+</sup> T cells analysed and CEACAM1- and TIM-3-expression assessed on the CFSE-positive cells.

**Cell lines.** 3T3 mouse fibroblast cells, the TK-1 mouse T cell lymphoma cell line, the HEK293T cell line, and CT26 mouse colorectal cancer cell lines were all purchased from ATCC as mycoplasma-free cell lines and culture conditions were all followed according to ATCC instructions. Human CEACAM1-4L transfected human Jurkat T-cell line and human CEACAM1-4L transfected HeLa cells have been previously described (provided by J. Shively and S. G. Owen)<sup>28,29</sup>.

**Antibodies and reagents.** An anti-CEACAM1 monoclonal antibody that binds to the C–C' loop of mCEACAM1 (ref. 30) (cc1, provided by K. V. Holmes and J. Shively) and isotype control (mouse IgG1) antibodies for the *in vivo* injection experiments were purchased from Bio X Cell. An Fc fusion protein consisting of the N-terminal domain of mouse CEACAM1 together with human IgG1, N-CEACAM1–Fc<sup>31</sup> was

provided by T. Gallagher. The following fluorochrome-conjugated antibodies were purchased from BioLegend: anti-IFN-γ (XMG1.2), anti-IL-2 (JES6-5H4), anti-IL-17A (TC11-18H10.1), anti-IL-10 (JES5-16E3) and PD-L1 (10F.9G2); and other antibodies were purchased from eBioscience: anti-CD3 (clone, 145-2C11), anti-CD28 (clone 37.51), anti-CD8α (clone, 53-6.7), anti-CD4 (clone, RM4-5), anti-TNF-α (clone, MP6-XT22) and anti-TIM-3 (clone, 8B.2C12). The BAT3 antibody (clone AB10517) was purchased from Millipore. Brefeldin A and monensin were from eBioscience. The anti-TIM-3 5D12, 2C12, 2E2, 2E12 and 3F9 antibodies were generated in V.K.K.'s laboratory and 5D12 was conjugated to phycoerythrin and allophycocyanin by BioLegend. The anti-human CEACAM1 antibodies specific for the N-terminal domain (34B1, 26H7 and 5F4) have been previously described and characterized in R.S.B.'s laboratory<sup>32,33</sup>. All the antibodies used for biochemical experiments if otherwise indicated, were purchased from Cell Signaling. Tunicamycin was purchased from Sigma.

**Primary human T-cell isolation.** Human peripheral blood mononuclear cells (PBMCs) were isolated from buffy coats from healthy anonymous donors according to an Institutional Review Board approved protocol from Brigham and Women's Hospital. CD4<sup>+</sup> T cells were purified from PBMC using CD4 Microbeads (Miltenyi) and cultured in RPMI-1640 supplemented with 10% FBS (Invitrogen), L-glutamine, non-essential amino acids, sodium pyruvate, 20 mM HEPES and 100 IU ml<sup>-1</sup> recombinant human IL-2 (R&D Systems). The T cells were stimulated with plate-bound anti-CD3 (OKT3) and anti-CD28 (CD28.2) antibodies (1 µg ml<sup>-1</sup> each, eBioscience).

**Cellular based studies on HIV-infected individuals.** In the analysis of HIV-infected individuals, informed consent was obtained in accordance with the guidelines for conduction of clinical research at the University of Toronto and Maple Leaf Clinic institutional ethics boards. Written informed consent was provided for this study, which was reviewed by research ethics board of the University of Toronto, Canada and of St. Michael's Hospital, Toronto, Canada. Individuals were recruited from a Toronto-based cohort (Maple Leaf Clinic and St. Michael's Hospital, Toronto, Canada). Samples were obtained from HIV-positive chronically infected antiretroviral treatment naive individuals (infected >1 year, with detectable viral load) and demographically matched HIV-seronegative individuals. Whole blood was collected in anti-coagulant treated tubes and PBMCs were isolated using Ficoll-Paque PLUS (GE Healthcare Bio-Sciences) and stored at –150 °C until future use. PBMCs from healthy HIV-1-uninfected and HIV-1-infected individuals were stained with CEACAM1 monoclonal antibody (clone 26H7), detected with a fluorophore-conjugated secondary antibody, followed by fluorophore-conjugated monoclonal antibodies to CD4, CD8 and CD3 (BioLegend), and TIM-3 (R & D Systems) to determine phenotype assessment. An Aqua amine dye (Invitrogen) was used to discriminate live and dead cells. For functional experiments, cells were stimulated after thawing with 1 µg ml<sup>-1</sup> of overlapping HIV-1 Clade B Gag peptide pool (National Institutes of Health AIDS Reagent Program), or 1 µg ml<sup>-1</sup> SEB (Sigma Aldrich) in the presence of brefeldin A (Sigma) for 6 h. Surface staining was followed by a fixation and permeabilization step. Intracellular staining for cytokines was performed using anti-IFN-γ (Biolegend). Cells were fixed in 2% formalin/PBS and acquired with a modified LSRII system (BD Biosciences). A total of 100,000 events were collected and analysed with FlowJo software (Tree Star).

**In situ proximity ligation assay.** Proximity ligation assay was performed using Duolink *in situ* PLA reagents on HEK293T cells co-transfected with human CEACAM1–Flag-tagged and human TIM-3–HA-tagged vectors. Primary antibodies used were anti-Flag or anti-HA antibodies and followed by secondary antibodies. Cells on slides were blocked with Duolink Blocking stock followed by the application of two PLA probes in 1× antibody diluent. The slides were washed in a wash buffer (1× TBS-T) for 5 min twice and processed for hybridization using Duolink Hybridization stock 1:5 in high purity water and followed by incubation for 15 min at 37 °C. Duolink Ligation was performed with ligase and the slides were incubated in a pre-heated humidity chamber for 15 min at 37 °C. Amplification was then achieved using Duolink Amplification stock containing the polymerase and the slides were incubated again in a pre-heated humidity chamber for 90 min at 37 °C. DNA was stained with DAPI. A proximity ligation assay probe generates a fluorescent signal only when it binds to two primary antibodies (anti-HA and anti-Flag) attached to two proteins that are in maximum distance of 30–40 nm (ref. 12) to each other. One individual dot represents the close proximity of two interacting proteins within the cells.

**Confocal microscopy of T–B-cell conjugates.** CD4<sup>+</sup> and CD8<sup>+</sup> T cells were isolated from human PBMCs by negative selection (EasySep, StemCell Technologies) and expanded by incubation with 1 µg ml<sup>-1</sup> anti-CD3 (clone OKT3; BioLegend), 1 µg ml<sup>-1</sup> anti-CD28 monoclonal antibody (clone 28.8; BioLegend) with 50 U ml<sup>-1</sup> of recombinant IL-2 (National Institutes of Health) in complete RPMI media for 5 days to upregulate CEACAM1 and TIM-3. After overnight rest, T cells were mixed in a 2:1 ratio with Cell Tracker Blue CMAC (Invitrogen)-labelled human EBV-transformed B cells pre-loaded with SEB (Sigma-Aldrich), and brought into contact by centrifugation at 200g for 5 min. Cells were incubated at 37 °C for 10 min and deposited onto poly-L-lysine coated coverslips for an additional 15 min. Cells were

fixed in 4% formaldehyde, permeabilized, blocked and stained with primary antibodies against CD3- $\epsilon$  (Abcam), CEACAM1 (clone 5F4) and TIM-3 (R&D Systems) and appropriate fluorescently labelled secondary antibody. After immunofluorescent labelling, coverslips were mounted on glass slides using ProLong Gold antifade reagent (Invitrogen) and cured in the dark at room temperature for 24 h. Imaging was performed on an Olympus IX81 confocal microscope. Z-stack images (0.15  $\mu$ m) of conjugates were acquired with a  $\times 60$  Plan-Apochromat oil objective, numerical aperture 1.42.

**Image processing and analysis.** All fluorescence images were background subtracted, annotated and exported using Olympus Fluoview FV100 image viewer. Colocalization analysis and Pearson's correlation coefficient was calculated from corresponding interface regions from two channels using Volocity. For presentation in figures, images were adjusted and cropped equally across related groups.

**Mutagenesis of human CEACAM1 and TIM-3.** Point mutations were introduced by PCR-based mutagenesis, using the QuikChange II Site-Directed Mutagenesis Kit (Agilent Technologies). The mutant oligonucleotides are listed in Extended Data Table 1a, b. Previously described vectors containing the human CEACAM1-3L variant<sup>34</sup> and human TIM-3 in the pDisplay vector (Invitrogen)<sup>35</sup> were used as the template for all mutations. PCR reactions for single amino acid mutations were run for 16 cycles of 30 s at 95 °C and 1 min at 55 °C, followed by 6 min at 68 °C. The resulting mutant plasmids were verified by Sanger DNA sequencing. TIM-3 amino acid residues were numbered according to National Center for Biotechnology Information database.

**Cell culture, transfection, tunicamycin treatment, immunoprecipitation and immunoblotting.** HEK293T cells transfected with the 1,200 ng of Flag-tagged human CEACAM1 wild-type or mutant vectors or 1,200 ng of Flag-tagged ITGA5 (NM\_002205; Origene) and 1,200 ng of HA-tagged human TIM-3 wild-type or mutant vectors or 1,200 ng of vector controls when mono-transfections were performed and cells transfected for 48 h. In some experiments, 6 h after transfection, transfected cells were treated with 2  $\mu$ g ml<sup>-1</sup> or 10  $\mu$ g ml<sup>-1</sup> tunicamycin provided in DMSO for the last 24 h of transfection. Transfected cells were washed once with cold PBS and lysed on ice with 0.5 ml of immunoprecipitation buffer containing 20 mM Tris-HCl, 0.15 M sodium chloride, pH 7.6, with protease inhibitor cocktail tablets (Roche) and 1.0% digitonin (Sigma). After 60 min, the cell lysates were spun at 14,000 r.p.m. for 30 min at 4 °C. The lysate was subsequently incubated with 5  $\mu$ l of protein A-Sepharose (Sigma) that had been pre-adsorbed with an equal volume of immunoprecipitation buffer followed by three 1-h incubations with protein A-Sepharose. The pre-cleared lysates were incubated for 2 h at 4 °C with 10  $\mu$ l of protein A-Sepharose beads bound to the specific antibodies. Some immunoprecipitates were incubated with agarose-HA antibody. The control and specific immunoprecipitates were washed with immunoprecipitation buffer and re-suspended in 30  $\mu$ l of Laemmli sample buffer without reducing agents. After boiling for 5 min, the proteins were resolved by SDS-PAGE in regular Tris-glycine buffer on a 4–20% Tris-Glycine Gel (Novex). The proteins were electrically transferred to a PVDF (polyvinylidene difluoride) membrane. After blocking with 5% skim milk in 0.05% PBS-Tween (PBS-T), the membranes were incubated for 12 h at 4 °C with primary antibodies. The membranes were further incubated with corresponding secondary antibodies for 1 h at room temperature and visualized by Amersham ECL Western Blotting Detection Reagents (GE Healthcare). The specific antibodies used for immunoprecipitation and immunoblotting in this assay were anti-Flag antibody produced in rabbit (Sigma), HA.11 (16B12) (Covance), anti-HA-Tag-agarose (MBL, Medical & Biological Laboratories Co), anti-HA antibody produced in rabbit and in mouse (Sigma) and anti-human CEACAM1 antibody, 5F4. Unsaturated films were digitally scanned and band intensities (densitometric analysis) were quantified using ImageJ (NIH).

**Co-immunoprecipitation of primary T cells.** Primary human CD4<sup>+</sup> T cells were purified and activated with plate bound anti-CD3 (1  $\mu$ g ml<sup>-1</sup>) and anti-CD28 (1  $\mu$ g ml<sup>-1</sup>) in the presence of recombinant (r) IL-2 (10 ng ml<sup>-1</sup>; NIH) for 5 days. Cells were then rested in rIL-2 containing (10 ng ml<sup>-1</sup>) complete medium for 3 days and re-stimulated with soluble anti-CD3 (1  $\mu$ g ml<sup>-1</sup>) for 2 days. Cell lysates were prepared as above and treated with N-glycanase by the manufacturer's suggested protocol (Promega). N-glycanase treated lysates were immunoprecipitated with anti-human TIM-3 antibodies (2E2, 2E12 and 3F9) or mouse IgG1 as control and the immunoprecipitates subjected to SDS-PAGE and transferred to PVDF membranes followed by immunoblotting with the mouse anti-human CEACAM1 monoclonal antibody, 5F4, as described above. Primary mouse splenocytes were prepared from *Ceacam1-4L*<sup>Tg</sup> *Ceacam1*<sup>-/-</sup> and *Ceacam1-4S*<sup>Tg</sup> *Ceacam1*<sup>-/-</sup> mice and cultured in the presence of anti-CD3 (1  $\mu$ g ml<sup>-1</sup>) or anti-CD3 (1  $\mu$ g ml<sup>-1</sup>) plus anti-CD28 (1  $\mu$ g ml<sup>-1</sup>) or medium only for 96 h. Cell lysates were prepared, immunoprecipitated with cc1 (anti-mCEACAM1) and immunoblotted with 5D12 (anti-mTIM-3) monoclonal antibodies as described above.

**Pulse-chase experiments, immunoprecipitation, endoglycosidaseH digestion and SDS-PAGE.** Cells were transfected as above with wild-type human HA-tagged

hTIM-3 or Thr101Ile variant of HA-tagged hTIM-3 with or without co-transfection with Flag-tagged human CEACAM1 vector. All transfections included a co-transfection with Flag-tagged influenza virus M2 protein (M2) encoding vector as a non-binding control. After 48 h, transfected HEK293T cells were incubated with methionine- and cysteine-free DMEM for 30 min at 37 °C. Cells were labelled with 10 mCi ml<sup>-1</sup> [<sup>35</sup>S]-methionine/cysteine (1,175 Ci mmol<sup>-1</sup>; PerkinElmer Life Sciences) at 37 °C for the indicated times and chased with DMEM at 37 °C for the indicated times. Cells were lysed in Nonidet P-40 lysis buffer (50 mM Tris, pH 7.4, 0.5% Nonidet P-40, 5 mM MgCl<sub>2</sub>, and 150 mM NaCl). Immunoprecipitations were performed using 30  $\mu$ l of anti-HA agarose (Roche) for 2 h at 4 °C with gentle agitation. For enzymatic digestions, immunoprecipitates were denatured in glycoprotein denaturing buffer (0.5% SDS, 1% 2-mercaptoethanol) at 95 °C for 5 min, followed by addition of sodium citrate (pH 5.5) to a final concentration of 50 mM, and incubated with endoglycosidaseH (New England Biolabs) at 37 °C for 2 h. Immune complexes were eluted by boiling in reducing sample buffer, subjected to SDS-PAGE (10%), and visualized by autoradiography. Densitometric quantification of radioactivity was performed on a PhosphorImager (Fujifilm BAS-2500) using Image Reader BAS-2500 V1.8 software (Fujifilm) and Multi Gauge V2.2 (Fujifilm) software for analysis. Quantification of human TIM-3 in pulse-chase metabolic labelling was calculated by the following formula representing the average of 2–4 data sets: (human TIM-3 signal intensity – M2 protein signal intensity)/M2 protein signal intensity).

**Transduction of primary human T cells with shRNA lentiviral vectors to silence human TIM-3.** Lentiviral vectors encoding short hairpin RNAs (shRNAs) (pLKO.1) targeting human *TIM3* and a control shRNA targeting *lacZ* were obtained from the Dana Farber DNA Resource Core. Lentiviral particles were produced as previously described<sup>36</sup>. Primary human CD4<sup>+</sup> T cells were transduced 3 days after stimulation and selected with 2  $\mu$ g ml<sup>-1</sup> puromycin (Sigma) for 5 days followed by gradual increases of puromycin (5  $\mu$ g ml<sup>-1</sup>) for the next 5 days and re-stimulated with anti-CD3 and CD28 for an additional 5 days. *TIM3* shRNA 1 (TRCN0000158033, 5'-C GTGGACCAAACTGAAGCTAT-3'); *TIM3* shRNA 2 (TRCN0000154618, 5'-G CACTGACCTTAAACAGGCAT-3'); *TIM3* shRNA 3 (TRCN0000157816, 5'-C AAATGCAGTAGCAGAGGAA-3'); *lacZ* control knockdown (TRCN0000072225; 5'-CTCTGGCTAACGGTACGCGTA-3').

**Transduction of primary mouse T cells with retroviral vectors.** For transduction of mouse primary CD4<sup>+</sup> T cells, previously described mouse TIM-3 and its mutant expressing retroviruses were prepared in HEK293T cells using *eco* and *gag/pol* viral envelope constructs (Clontech Inc.)<sup>17,37,38</sup>. Viral supernatants were collected 48 h after transfection and centrifuged and filtered with a 0.45- $\mu$ m filters. CD4<sup>+</sup> T cells were prepared at 10<sup>6</sup> cells per ml in the presence of polybrene (Sigma; 8 mg ml<sup>-1</sup>). Viral particles and CD4<sup>+</sup> T cells were co-incubated by spin inoculation at 2,000g for 60 min on two sequential days. To detect ectopic expression of the constructs in T cells, cells were allowed to rest for 3 days in 2 ng ml<sup>-1</sup> IL-2 and sorted to obtain cells with the highest (45%) GFP expression. Sorted cells were titrated into *in vitro* T-cell assays.

**Galectin-9 induction of apoptosis.** Naive CD4<sup>+</sup> T cells were polarized under T<sub>H</sub>1 inducing conditions. In brief, naive CD4<sup>+</sup> T cells (2  $\times$  10<sup>6</sup> per ml) were stimulated with plate-bound anti-CD3 specific antibodies (10  $\mu$ g ml<sup>-1</sup>) for 48 h. Soluble anti-CD28 antibodies (1  $\mu$ g ml<sup>-1</sup>), recombinant (r) IL-2 (10 ng ml<sup>-1</sup>, NIH) and blocking antibodies to IL-4 (10  $\mu$ g ml<sup>-1</sup>) and rIL-12 (5 ng ml<sup>-1</sup>) were added and the cells cultured for 7 days. Cells were washed and treated with galectin-9 (2  $\mu$ g ml<sup>-1</sup>; eBioscience) for 8 h to induce apoptosis. Apoptosis was detected by flow cytometry using apoptosis staining kit (Roche) that stains for annexin V and propidium iodide.

**Protein purification, crystallization, X-ray data collection and model determination.** Competent *E. coli* BL21 DE3 cells were transformed with a pET9a vector carrying a gene insert coding for a single-chain of a protein consisting of the human CEACAM1 IgV domain (residues 1–107), a 20 amino acid linker (4 repeats of a Gly-Gly-Gly-Gly-Ser motif), and the human IgV domain of TIM-3 (residues 1–105) with a hexa-histidine tag appended to the C terminus. To express the protein, transformants were grown in 1 litre of LB broth under antibiotic selection and induced with 0.1 mM isopropyl- $\beta$ -D-thiogalactoside (IPTG) after reaching an absorbance at 600 nm (*A*<sub>600 nm</sub>) of approximately 0.8. Cells were grown for an additional 16 h at 18 °C and collected by centrifugation. Protein was purified from inclusion bodies, which were solubilized in a chaotropic buffer and refolded in a buffer containing 200 mM Tris, pH 8.5, 0.4 M arginine-HCl, 2 mM EDTA, 5 mM cysteamine and 0.5 mM cystamine, as reported previously<sup>39</sup>. Refolded protein was loaded onto a Ni-NTA affinity column and washed several times with a buffer containing 50 mM HEPES, pH 7.5, 300 mM NaCl, 2.5 mM CaCl<sub>2</sub>, 10% glycerol and 10–30 mM imidazole. The protein was eluted with 50 mM HEPES buffer, pH 7.2, 300 mM NaCl, 2.5 mM CaCl<sub>2</sub>, 10% glycerol and 300 mM imidazole. Eluted protein fractions were analysed by SDS-PAGE, pooled, concentrated and further purified with a Superdex 75 gel filtration column (GE Healthcare) in a buffer containing 50 mM HEPES buffer, pH 7.5, 200 mM NaCl, 2.5 mM CaCl<sub>2</sub> and 10% glycerol. Circular dichroism



was conducted to confirm proper protein folding. Purified protein (purity >95%, estimated by SDS–PAGE) was concentrated to 2.5 mg ml<sup>−1</sup> for crystallization.

The initial crystal growth trials were performed by vapour diffusion using a Phoenix robotic system (Art Robbins) with Index Screens 1 and 2 (Hampton Research) and optimized using an additive screen (Hampton Research). Optimized crystals were formed in 4 µl drops (2 µl protein, 2 µl mother liquor), equilibrated against a mother liquor containing 100 mM Tris 8.0 and 25% PEG 3350, at room temperature in a sitting drop setup. For data collection, crystals were cryoprotected in solution containing 25% PEG 3350 and 12% glycerol. X-ray data were collected using the X25 beamline at the National Synchrotron Light Source (NSLS). The data were processed with Mosflm<sup>40</sup> and the CCP4 software suite<sup>41</sup>.

A crystallographic model of the single-chain human CEACAM1–TIM-3 protein was built by a molecular replacement strategy using a search model made by modifying published structures of the IgV (N) domain of human CEACAM1 (Protein Data Bank (PDB) code 2GK2) and mouse TIM-3 (PDB code 2OYP). Specifically, all residues were changed to alanine to reduce model bias and all small molecule ligands, water and metals were removed. Molecular replacement was performed with the program Molrep. Model building and structural refinement were performed with the PHENIX software<sup>42</sup> and COOT<sup>43</sup>, respectively. Five per cent of the total reflection data were excluded from the refinement cycles and used to calculate the free *R* factor (*R*<sub>free</sub>) for monitoring refinement progress. Repeated rounds of model building and refinement with group B-factors refinement strategy and Torsion-NCS restraints resulted in a final crystallographic *R*<sub>work</sub>/*R*<sub>free</sub> of 34.5%/37.3% using all data to 3.4 Å resolution. The X-ray data and refinement statistics are shown in Extended Data Table 2. Most of the residues were well defined and verified by PDB validation report (data not shown). However, the β-strands associated with the hTIM-3 AB–ED (ref. 13) face did not show the secondary structure of β-sheet formation, and two disulphide bonds (Cys 32–Cys 110 and Cys 53–Cys 62) observed in the mouse TIM-3 structure were missing from the resolved structure. A 2*F*<sub>o</sub> − *F*<sub>c</sub> electron density map of the human CEACAM1 (IgV)–TIM-3 (IgV) single chain was calculated using PHENIX. All the figures were drawn using PyMOL (The PyMOL Molecular Graphics System, Schrödinger, LLC) and labels were added using Adobe Photoshop.

**Protein purification, crystallization, data collection and structure determination of CEACAM1 (IgV).** Competent *E. coli* BL21 DE3 cells were transformed with a pET9a vector carrying the human CEACAM1 IgV domain gene insert with a N-terminal GST tag. To express the protein, transformants were grown in 1 litre of LB broth under antibiotic selection and induced with 0.1 mM IPTG after reaching an *A*<sub>600 nm</sub> of approximately 0.8. Cells were grown for an additional 16 h at 18 °C and collected by centrifugation. Cell pellets were suspended in 20 mM Tris–HCl buffer, pH 7.5, containing 150 mM NaCl and 10% glycerol, and lysed by sonication. After centrifugation, supernatant was loaded onto a GST column (GE Healthcare) and washed several times with 50 mM Tris–HCl buffer, pH 7.5, containing 150 mM NaCl, 2.5 mM CaCl<sub>2</sub>, 0.1% Triton X-100, and 10% glycerol. After on-column removal of the GST affinity tag with thrombin, protein was eluted with 25 mM Tris–HCl buffer, pH 7.5, containing 200 mM NaCl, 2.5 mM CaCl<sub>2</sub> and 10% glycerol. Eluted protein fractions, as judged by SDS–PAGE, were concentrated and purified through a Superdex 75 gel filtration column (GE Healthcare) by FPLC system with a buffer containing 50 mM Tris–HCl buffer, pH 7.5, containing 200 mM NaCl, 2.5 mM CaCl<sub>2</sub> and 10% glycerol. A single peak for human CEACAM1 IgV domain was collected and the protein was concentrated to 5 mg ml<sup>−1</sup> for crystallization. A purity of >95%, was verified by SDS–PAGE.

The initial crystal growth trials were performed in a 96-well format using a Phoenix robotic system (Art Robbins) with Index Screens 1 and 2 (Hampton Research) at room temperature. Optimized crystals were formed in well solution containing 60% Tascimate, pH 8, and 1% β-octyl glucoside at room temperature. For data collection, crystals were cryoprotected in solution containing 60% Tascimate, pH 8 and 18% glycerol. X-ray data were collected using beamline X6A at the National Synchrotron Light Source (NSLS). The data were processed with Mosflm<sup>40</sup> and the CCP4 software suite<sup>41</sup>. The structure of the human CEACAM1 N (IgV) domain was determined by the molecular replacement method using a modified N-terminal domain of human CEACAM1 as a search model (PDB code 2GK2) as the starting model, with all residues changed to alanine to reduce model bias and all small molecule ligands, water and metals were removed. Molecular replacement was performed with Molrep, refinements were performed with PHENIX<sup>42</sup> and intermittent model building was performed with COOT<sup>43</sup>, as described above. Subsequent refinements cycles and model building with individual B-factors and Torsion-NCS restraints strategy led to the final crystallographic *R*<sub>work</sub>/*R*<sub>free</sub> of 20.3%/24.2% using all data to 2.0 Å resolution. Electron densities were also identified for bound molecules of beta-octyl glucoside and malonic acid. The X-ray data and structure refinement statistics are shown in Extended Data Table 2. All residues were well defined and verified by PDB validation report (data not shown). A 2*F*<sub>o</sub> − *F*<sub>c</sub> electron density map was calculated using PHENIX<sup>42</sup>.

**Structural modelling of CEACAM1–TIM-3 interactions.** To model potential interactions between CEACAM1 and TIM-3, superimposition data as shown in Extended Data Fig. 2d, e, of previously described immunoglobulin-variable (IgV) domains of mouse CEACAM1<sup>14</sup> (PDB code 1L62) and mouse TIM-3 (PDB 2OYP)<sup>13</sup> was generated in Pymol (The PyMOL Molecular Graphics System, Schrödinger, LLC), which revealed structural similarity as shown by a score of 2.42 by root mean square deviation (r.m.s.d.) analysis as calculated by Pymol<sup>44</sup>. In addition, multiple sequence alignments were performed using ClustalW<sup>45</sup>. From this analysis, we considered mouse CEACAM1 and TIM-3 as structurally similar proteins. Furthermore, as previous structural data for mouse CEACAM1 supported the formation of CEACAM1 homodimers between its N-terminal domains<sup>14</sup>, and given our evidence that mouse CEACAM1 and TIM-3 are highly similar, we further reasoned that mouse TIM-3 and its human orthologue might heterodimerize with CEACAM1 through their membrane-distal N-terminal domains. This hypothesis was modelled by us in the following manner. We aligned CEACAM1 symmetrically for *cis* modelling and asymmetrically for *trans* modelling along the five amino acid residues within the FG–CC' loops of mTIM-3 as reported previously<sup>13</sup> that describe putative docking sites for an unknown ligand for mouse TIM-3. The ROSIE Docking Server was used to analyse the docking (<http://rosettadock.graylab.jhu.edu/documentation/docking>)<sup>46,47</sup>. We identified the ten best scoring structures from the run in rank order by energy. The top two in *cis* and the best in *trans* are shown in Extended Data Fig. 2g. The score (y axis) indicates the energy of the models (the lower the better). The r.m.s. (x axis) indicates the RMS values between the input and output docking models. The docking models identified were further assessed in PyMOL to determine whether the models identified predicted hydrogen bonding within the putative mouse CEACAM1–TIM-3 interface. On the basis of this we identified potential amino acids that may have an important role in mouse CEACAM1–TIM-3 binding, which is summarized in the table associated with Extended Data Fig. 2g.

We used the amino acids of mouse CEACAM1 and mouse TIM-3 modelled at the interface between these two proteins to predict potential contact sites for the human orthologues of CEACAM1 and TIM-3. Furthermore, we searched human exomic databases for allelic variants in these predicted human amino acids. Finally, we used all of this information to make structural modifications of human CEACAM1 and TIM-3 by site-directed mutagenesis to biochemically examine our hypothesis and confirm this by X-ray crystallographic analysis of a human CEACAM1–TIM-3 single-chain protein.

**IBD case-control samples and genotyping.** German Crohn's disease and ulcerative colitis patients were recruited at the Department of General Internal Medicine of the Christian-Albrechts-University Kiel and Charité Universitätsmedizin Berlin, or nationwide with the support of the German Crohn and Colitis Foundation and the Bundesministerium für Bildung und Forschung (BMBF) competence network 'IBD' (Extended Data Table 3a). The patients were classified according to clinical, radiological, histological and endoscopic (that is, type and distribution of lesions) according to accepted criteria<sup>48,49</sup>. Healthy control individuals were obtained from the popgen biobank<sup>50</sup>. All controls were drawn from a population-representative sample. Given the low prevalence of both Crohn's disease and ulcerative colitis and the fact that all control individuals self-reported to have neither Crohn's disease nor ulcerative colitis, control individuals were designated 'healthy'. Written, informed consent was obtained from all study participants and all protocols were approved by the institutional ethical review committees of the participating centres. The variants described in Extended Data Table 3b were genotyped using the Sequenom iPLEX and the Life Technologies Taqman system. rs147827860, the single nucleotide polymorphism encoding the hypomorphic Thr101Ile allele of *TIM3*, is a rare variant with a minor allele frequency of 0.004 in individuals of European descent predicting homozygous carriage with a frequency of 0.000016 (or 1:62,500). None of the publicly accessible databases lists any homozygous occurrences of this variant. Association in the case-control data was tested with Haploview 4.2 (ref. 51). An exact test for Hardy–Weinberg equilibrium was performed using the DeFinetti program (Strom, T. M. & Wienker, T. F., <http://ihg.gsf.de/cgi-bin/hw/hwa1.pl>).

**Purification of IgV domain of human TIM-3.** The TIM-3 IgV domain was expressed with a C-terminal GST tag as inclusion bodies from pET9a in *E. coli* BL21 DE3 cells. Inclusion bodies were refolded as described previously in refolding buffer containing 200 mM Tris, pH 8.5, 0.4 M arginine–HCl, 2 mM EDTA, 5 mM cysteamine and 0.5 mM cystamine<sup>39</sup>. Refolded proteins were purified by affinity and gel filtration chromatography. The proteins were maintained in a solution containing 25 mM HEPES buffer, pH 7.2, containing 200 mM NaCl, 2.5 mM CaCl<sub>2</sub> and 5% glycerol. Circular dichroism was conducted to ascertain proper protein folding. A purity of >95%, was verified by SDS–PAGE.

**GST protein pull-down assays.** Glutathione-S-transferase (GST) pull-down analyses<sup>52</sup> of the interaction between IgV domains of human TIM-3 and human CEACAM1 were performed in the following manner. The GST–hTIM-3 protein was stored in refolding buffer, and buffer-exchanged with PBS before performing the assay. The



IgV domain of GST-hTIM-3 proteins were coupled to the glutathione-agarose column following the manufacturer's instructions (Thermo scientific) and in-column incubated with HEK293T cell lysates derived from cells transfected with Flag-tagged human CEACAM1. The columns were washed extensively and bound proteins were eluted by glutathione. Eluted protein complexes were immunoblotted for the detection of CEACAM1 (immunoblot with anti-Flag) and TIM-3 (immunoblot with anti-GST).

**TIM-3 ligand precipitation assays.** Extracellular membrane-associated proteins on live TK-1 cells were labelled with biotin (EZ-Link Sulfo-NHS-LC-Biotin; Pierce). Whole-cell lysates were prepared and were incubated with 5 µg human IgG, full length TIM-3-Ig or soluble (s)TIM-3-Ig in the presence of protein G-agarose beads (Roche). Beads were washed and boiled with 1× SDS-PAGE loading buffer. Supernatants were collected by centrifugation, and half of each was digested with PNGase F (New England Biolabs). Samples were separated by SDS-PAGE and the TIM-3-Ig fusion protein-specific binding proteins were detected by immunoblot or silver staining.

**Mouse model of IBD<sup>53</sup>.** Splenic mononuclear cells were obtained from female wild-type, *Ceacam1*<sup>−/−</sup>, *Ceacam1*<sup>4L</sup> *Ceacam1*<sup>−/−</sup>, *Tim3*<sup>Tg</sup> *Ceacam1*<sup>−/−</sup>, *Tim3*<sup>Tg</sup> *Ceacam1*<sup>+/+</sup> (*Tim3*<sup>Tg</sup>/wild-type) mice and CD4<sup>+</sup> T cells were isolated using anti-CD4 (L3T4) MACS magnetic beads (Miltenyi Biotec) according to the manufacturer's instructions. Enriched CD4<sup>+</sup> T cells (94–97% pure as estimated by FACS) were then labelled with PerCP-conjugated anti-CD4, allophycocyanin-conjugated anti-CD44, and FITC-conjugated anti-CD62L and phycoerythrin-conjugated anti-TIM-3 (5D12 or 2C12). Subpopulations of CD4<sup>+</sup> T cells were identified by three-colour sorting on a FACS Aria (Becton Dickinson). All populations were defined as naive and 98.0% pure on re-analysis. To induce chronic colitis in animals, 5 × 10<sup>5</sup> CD4<sup>+</sup> CD44<sup>lo</sup> CD62L<sup>high</sup> (naive) T cells were adoptively transferred intraperitoneally into 6–10-week-old *Ceacam1*<sup>−/−</sup> *Rag2*<sup>−/−</sup> recipient mice. Weights were measured every week, and mice were euthanized by CO<sub>2</sub> for histological evaluation of colitis and lamina propria mononuclear cells. Owing to the severity of colitis associated with adoptive transfer of naive *Ceacam1*<sup>−/−</sup> CD4<sup>+</sup> T cells, significant mortality was observed resulting in unequal group size at the time of culling necessitating post-hoc correction in the statistical analysis (see 'statistical methods').

**Colorectal carcinogenesis models.** To examine the effect of CEACAM1 and TIM-3 on tumour incidence and multiplicity in an inflammation induced colorectal cancer model, mice were intraperitoneally injected at 6 weeks of age with the carcinogen AOM (Sigma-Aldrich) at 10 mg kg<sup>−1</sup> body weight as previously reported<sup>54,55</sup>. One week later, the mice were started on the first of two 21-day DSS cycles, consisting of a period of 7 days with the tumour promoter DSS at 2.5% or 1.5% in the drinking water followed by 7 days of receiving regular water. Mice showing signs of morbidity were culled. Colons were removed and flushed with PBS buffer and cut longitudinally. Colon tissue sections were either paraffin-embedded for immunohistochemistry or analysis of lamina propria mononuclear cells associated with the tumours. CT26 colorectal carcinoma cells were diluted into 10 × 10<sup>6</sup> cells per ml of PBS. After shaving, either wild-type or *Ceacam1*<sup>−/−</sup> mice received 200 µl of the CT26 cell suspension subcutaneously in the left flank. All antibody treatments were applied intraperitoneally into the right flank according to the schedules described in Extended Data. Two-hundred micrograms of each antibody or its isotype control were administered according to the schedules described (CEACAM1, cc1 clone; TIM-3 5D12 clone; PDL1, 10F.9G2). Tumour growth was assessed three times per week at 7 days after tumour inoculation by measuring the width and length of the tumours (mm) using calipers and the areas defined (mm<sup>2</sup>; width × length). Mice were euthanized on the basis of the extent of tumour growth observed in the isotype antibody treated control animals to minimize animal suffering in accordance with the approved animal protocol.

**Histopathological examination of colitis and colitis-associated colorectal cancer.** Colons were removed from mice after termination and dissected free from the anus to the caecum. Colonic contents were removed and colons cleaned with PBS before fixation in 4% paraformaldehyde or 10% neutral buffered formalin followed by routine paraffin embedding. After paraffin embedding, 0.5-mm sections were cut and stained with haematoxylin and eosin. Sections were examined and colitis was scored in a blinded fashion (with respect to genotype and experimental protocol) by one of the authors (J.N.G.). Each of four histological parameters for the severity of colitis was scored as absent (0), mild to severe (1–6): mononuclear cell infiltration, polymorphonuclear cell infiltration, epithelial hyperplasia, epithelial injury and extent of inflammation, modified from a previous study<sup>56</sup>. For AOM/DSS, colitis-associated colorectal cancer studies, an expert pathologist (J.N.G.) examined all tissues for dysplasia, adenoma, low-to-high-grade adenoma and adenocarcinoma in a blinded fashion.

**RNA, complementary DNA and quantitative PCR.** RNA was isolated from whole cells using the Qiagen microRNA extraction kit following the manufacturer's instructions. RNA was quantified spectrophotometrically, and complementary DNA was reverse-transcribed using the cDNA archival kit (Applied Biosystems) following the manufacturer's guidelines. All primers were obtained from Applied Biosystem. The cDNA samples were subjected to 40 cycles of amplification in an ABI Prism

7900 Sequence Detection System instrument according to the manufacturer's protocol. Quantification of relative messenger RNA expression was determined by the comparative C<sub>t</sub> (critical threshold) method as described whereby the amount of target mRNA, normalized to endogenous cycles of amplification was determined by the formula 2<sup>−ΔΔC<sub>t</sub></sup>.

**Gene expression analyses by NanoString.** Naive T cells from wild-type, *Tim3*<sup>Tg</sup> and *Tim3*<sup>Tg</sup> *Ceacam1*<sup>−/−</sup> mice were transferred into *Ceacam1*<sup>−/−</sup> *Rag2*<sup>−/−</sup> recipients to induce colitis. For the analyses of transcriptional regulation, colonic infiltrating lamina propria mononuclear cells were isolated at 6 weeks after T-cell transfer and immediately lysed in RLT buffer (Qiagen). For RNA analysis, 100 ng of total RNA was used and hybridized to a custom designed gene CodeSet (non-enzymatic RNA profiling using bar-coded fluorescent probes) according to the manufacturers protocol (Nanostring Technologies). Barcodes were counted (1,150 fields of view per sample) on an nCounter Digital Analyzer following the manufacturer's protocol (NanoString Technologies, Inc.).

**Cytokine measurements.** To measure cytokine production by ELISA, 1 × 10<sup>5</sup> CD4<sup>+</sup> T cells were cultured in 200 µl RPMI 1640 (Sigma-Aldrich) supplemented with 10% heat-inactivated FBS, 500 U ml<sup>−1</sup> penicillin, 100 µg ml<sup>−1</sup> streptomycin (Sigma-Aldrich), 10 mM HEPES, 1% nonessential amino acids and 50 µM 2-mercaptoethanol (Life Technologies Invitrogen), which was termed complete RPMI 1640, in the presence of 5 µg ml<sup>−1</sup> plate-bound anti-CD3 and 2 µg ml<sup>−1</sup> soluble anti-CD28 monoclonal antibodies on flat-bottom 96-well plates (Costar), at 37 °C in a humidified atmosphere incubator containing 5% CO<sub>2</sub> for 48 h or otherwise indicated. Culture supernatants were removed and analysed for the production of cytokines such as IFN-γ, IL-2 or TNF-α. Cytokine concentrations were determined using specific ELISAs (R&D Systems) according to the manufacturer's recommendations. To measure cytokine production by intracellular cytokine staining on colitis-associated CD4<sup>+</sup> T cells obtained from the dissection of mesenteric lymph nodes, lymph nodes or lamina propria were incubated at 37 °C for 4–6 h in complete RPMI 1640 medium with 50 ng ml<sup>−1</sup> PMA (Sigma-Aldrich), 500 ng ml<sup>−1</sup> ionomycin (Sigma-Aldrich), and 1 µl ml<sup>−1</sup> GolgiPlug (BD). Surface staining was performed for 30 min, after which the cells were resuspended in fixation/permeabilization solution (Cytofix/Cytoperm kit; BD). Intracellular cytokine staining using antibodies as described earlier was performed according to the manufacturer's instructions. To measure cytokine production by intracellular cytokine staining in draining lymph node CD4<sup>+</sup> or CD8<sup>+</sup> T cells derived from CT26-tumour bearing mice, cells were stimulated with 5 µg ml<sup>−1</sup> AH1-peptide (SPSYVYHQF) or tumour-infiltrating lymphocytes (TIL) with 1 µg ml<sup>−1</sup> plate-bound monoclonal antibody (mouse anti-CD3 antibody for 48 h). The cells were counterstained with monoclonal antibody against CD8-FITC with or without AH1-tetramer (Medical Biological Laboratories) which detects the MuLV gp70 peptide (SPSYVYHQF) in the context of H-2L<sup>d</sup> and analysed by flow cytometry.

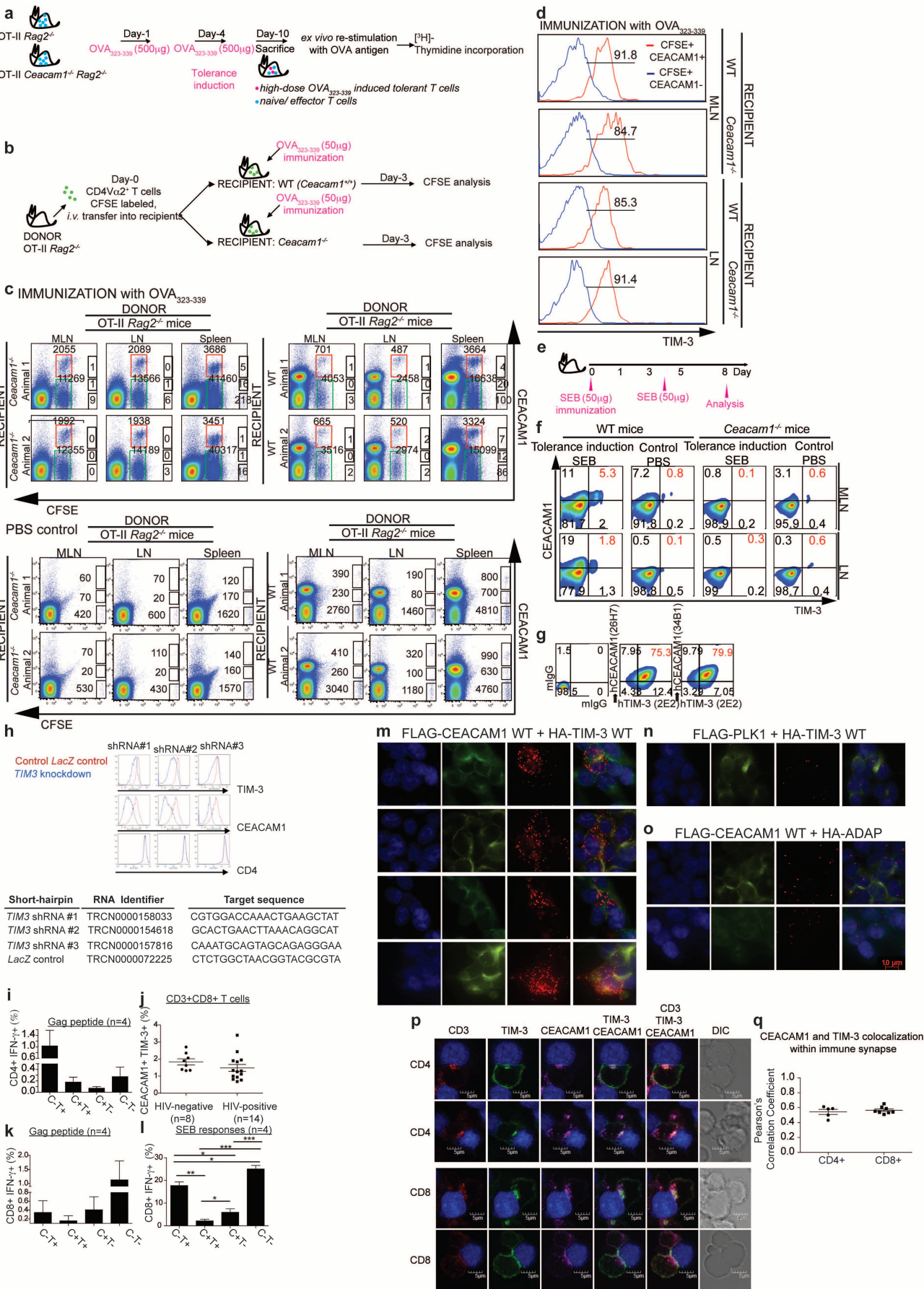
**Surface plasmon resonance.** Studies were performed on Biacore 3000 (GE Healthcare). GST-hTIM-3 protein, or GST alone as control, was immobilized on a CM5 sensor chip using amine coupling chemistry as per the manufacturer's instructions. The coupling was performed by injecting 30 µg ml<sup>−1</sup> of protein into 10 mM sodium acetate, pH 5.0 (GE Healthcare). HBS-EP buffer (0.01 M HEPES, 0.15 M NaCl, 3 mM EDTA, 0.005% surfactant P20; GE Healthcare) was used as running buffer and dilution buffer. Serial dilutions of analytes in HBS-EP were injected at 25 °C with a 25 µl min<sup>−1</sup> flow rate and data collected over time. The surface was regenerated between different dilutions with 10 mM glycine-HCl, pH 2.5 (GE Healthcare). For blockade of single-chain binding to GST-hTIM-3, single-chain hCEACAM1-hTIM-3 protein was injected alone or together with antibody (anti-human CEACAM1 monoclonal antibody, 26H7) or TIM-3-specific peptide (residues 58–77, 58-CPV FECGNVLTDERDVNY-77) with IgG1 mouse antibody or scrambled peptide (TLCVCFVNPYDVRVNDEREG) used as controls, respectively. All data were zero adjusted, and the reference cell value was subtracted.

**Isolation of lamina propria mononuclear cells and TILs.** Colonic lamina propria mononuclear cells, colonic polyp-derived TILs and subcutaneous CT26 derived TILs were isolated as follows. Total colons from each group were longitudinally cut and washed with HBSS (free of Ca<sup>2+</sup> and Mg<sup>2+</sup>) to remove faeces and debris. Subcutaneous CT26 tumours were dissected and dissociated in a gentle MACS dissociator (Miltenyi Biotec). Tumours, tumour-free colon pieces and CT26 tumour pieces were further finely minced and incubated in HBSS containing 5 mM EDTA, 0.145 mg ml<sup>−1</sup> dithiothreitol, 1 M HEPES, 10% FBS, and 1% penicillin/streptomycin at 37 °C for 15 min for two cycles to dissociate the epithelial monolayers. EDTA was then removed by three washes in HBSS. The colon tissue and tumour specimens then were digested in RPMI 1640 containing 0.4 mg ml<sup>−1</sup> collagenase D (Roche) and 0.01 mg ml<sup>−1</sup> DNase I (Roche) for 20 min (tumour-free colon pieces) or 30 min (tumour pieces) at 37 °C on a shaking platform and further digested with collagenase IV (Roche) at 0.01 mg ml<sup>−1</sup> DNase I (Roche) for 20 min (tumour-free colon pieces) or 30 min (tumour pieces) at 37 °C on a shaking platform. After enzymatic treatment, digested tissue was passed through a 70-µm cell strainer and the flow-through

medium containing the mononuclear cells was collected and centrifuged at 400g for 10 min and pelleted and subjected to a 40:80% Percoll (GE Healthcare) gradient followed by centrifugation for 20 min at 2,500 r.p.m. at room temperature. Lamina propria mononuclear cells or TILs were collected at the interphase of the gradient, washed once, and resuspended in RPMI 1640 complete medium for further analyses.

**Statistical methods.** Standard two-tailed *t*-test was applied throughout except for the following exceptions. A Welch *t*-test was applied in the setting of unequal variance. The Mann–Whitney *U* test was applied when data were demonstrated to not follow a Gaussian distribution. In experiments where more than two groups were compared, one-way analysis of variance (ANOVA) was performed followed by application of post-hoc correction using Bonferroni's multiple comparison test or Dunnett's correction followed by Friedman test. Kruskal–Wallis with post-hoc Dunn's correction was used in making multiple comparisons of groups with unequal size. Comparisons of mortality were made by Kaplan–Meier survival curve analysis with log-rank test to assess differences in cancer survival. Pearson's coefficient for correlation (*r*) test was performed for comparison of variables. GraphPad Prism version 5.0b was used for calculations. *P* values of 0.05 were considered significant.

24. Leung, N. *et al.* Deletion of the carcinoembryonic antigen-related cell adhesion molecule 1 (*Ceacam1*) gene contributes to colon tumor progression in a murine model of carcinogenesis. *Oncogene* **25**, 5527–5536 (2006).
25. Bansal-Pakala, P., Jember, A. G. & Croft, M. Signaling through OX40 (CD134) breaks peripheral T-cell tolerance. *Nature Med.* **7**, 907–912 (2001).
26. Jeon, M.-S. *et al.* Essential role of the E3 ubiquitin ligase Cbl-b in T cell energy induction. *Immunity* **21**, 167–177 (2004).
27. Moon, J. J. *et al.* Tracking epitope-specific T cells. *Nature Protocols* **4**, 565–581 (2009).
28. Chen, C.-J. & Shively, J. E. The cell-cell adhesion molecule carcinoembryonic antigen-related cellular adhesion molecule 1 inhibits IL-2 production and proliferation in human T cells by association with Src homology protein-1 and down-regulates IL-2 receptor. *J. Immunol.* **172**, 3544–3552 (2004).
29. Patel, P. C. *et al.* Inside-out signaling promotes dynamic changes in the carcinoembryonic antigen-related cellular adhesion molecule 1 (CEACAM1) oligomeric state to control its cell adhesion properties. *J. Biol. Chem.* **288**, 29654–29669 (2013).
30. Smith, A. L. *et al.* Monoclonal antibody to the receptor for murine coronavirus MHV-A59 inhibits viral replication in vivo. *J. Infect. Dis.* **163**, 879–882 (1991).
31. Gallagher, T. M. A role for naturally occurring variation of the murine coronavirus spike protein in stabilizing association with the cellular receptor. *J. Virol.* **71**, 3129–3137 (1997).
32. Morales, V. M. *et al.* Regulation of human intestinal intraepithelial lymphocyte cytolytic function by biliary glycoprotein (CD66a). *J. Immunol.* **163**, 1363–1370 (1999).
33. Watt, S. M. Homophilic adhesion of human CEACAM1 involves N-terminal domain interactions: structural analysis of the binding site. *Blood* **98**, 1469–1479 (2001).
34. Chen, D. *et al.* Carcinoembryonic antigen-related cellular adhesion molecule 1 isoforms alternatively inhibit and costimulate human T cell function. *J. Immunol.* **172**, 3535–3543 (2004).
35. Hastings, W. D. *et al.* TIM-3 is expressed on activated human CD4<sup>+</sup> T cells and regulates Th1 and Th17 cytokines. *Eur. J. Immunol.* **39**, 2492–2501 (2009).
36. Pertel, T. *et al.* TRIM5 is an innate immune sensor for the retrovirus capsid lattice. *Nature* **472**, 361–365 (2011).
37. Holst, J., Vignali, K. M., Burton, A. R. & Vignali, D. A. A. Rapid analysis of T-cell selection *in vivo* using T cell-receptor retrogenic mice. *Nature Methods* **3**, 191–197 (2006).
38. Persons, D. A. *et al.* Retroviral-mediated transfer of the green fluorescent protein gene into murine hematopoietic cells facilitates scoring and selection of transduced progenitors *in vitro* and identification of genetically modified cells *in vivo*. *Blood* **90**, 1777–1786 (1997).
39. Zhang, X., Schwartz, J.-C. D., Almo, S. C. & Nathenson, S. G. Expression, refolding, purification, molecular characterization, crystallization, and preliminary X-ray analysis of the receptor binding domain of human B7-2. *Protein Expr. Purif.* **25**, 105–113 (2002).
40. Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D* **67**, 271–281 (2011).
41. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
42. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
43. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
44. Mizuguchi, K. & Go, N. Seeking significance in three-dimensional protein structure comparisons. *Curr. Opin. Struct. Biol.* **5**, 377–382 (1995).
45. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics* **Chapter 2**, Unit 2.3 (2002).
46. Lyskov, S. & Gray, J. J. The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.* **36**, W233–W238 (2008).
47. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: what the ROSETTA protein modeling suite can do for you. *Biochemistry* **49**, 2987–2998 (2010).
48. Lennard-Jones, J. E. Classification of inflammatory bowel disease. *Scand. J. Gastroenterol. Suppl.* **170**, 2–6 (1989).
49. Truelove, S. C. & Pena, A. S. Course and prognosis of Crohn's disease. *Gut* **17**, 192–201 (1976).
50. Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet.* **9**, 55–61 (2006).
51. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
52. Ye, K. *et al.* Phospholipase C $\gamma$ 1 is a physiological guanine nucleotide exchange factor for the nuclear GTPase PIKE. *Nature* **415**, 541–544 (2002).
53. Powrie, F., Leach, M. W., Mauze, S., Caddle, L. B. & Coffman, R. L. Phenotypically distinct subsets of CD4<sup>+</sup> T cells induce or protect from chronic intestinal inflammation in C. B-17 scid mice. *Int. Immunol.* **5**, 1461–1471 (1993).
54. Okayasu, I., Ohkusa, T., Kajiura, K., Kanno, J. & Sakamoto, S. Promotion of colorectal neoplasia in experimental murine ulcerative colitis. *Gut* **39**, 87–92 (1996).
55. Neufert, C., Becker, C. & Neurath, M. F. An inducible mouse model of colon carcinogenesis for the analysis of sporadic and inflammation-driven tumor progression. *Nature Protocols* **2**, 1998–2004 (2007).
56. Adolph, T. E. *et al.* Paneth cells as a site of origin for intestinal inflammation. *Nature* **503**, 272–276 (2013).

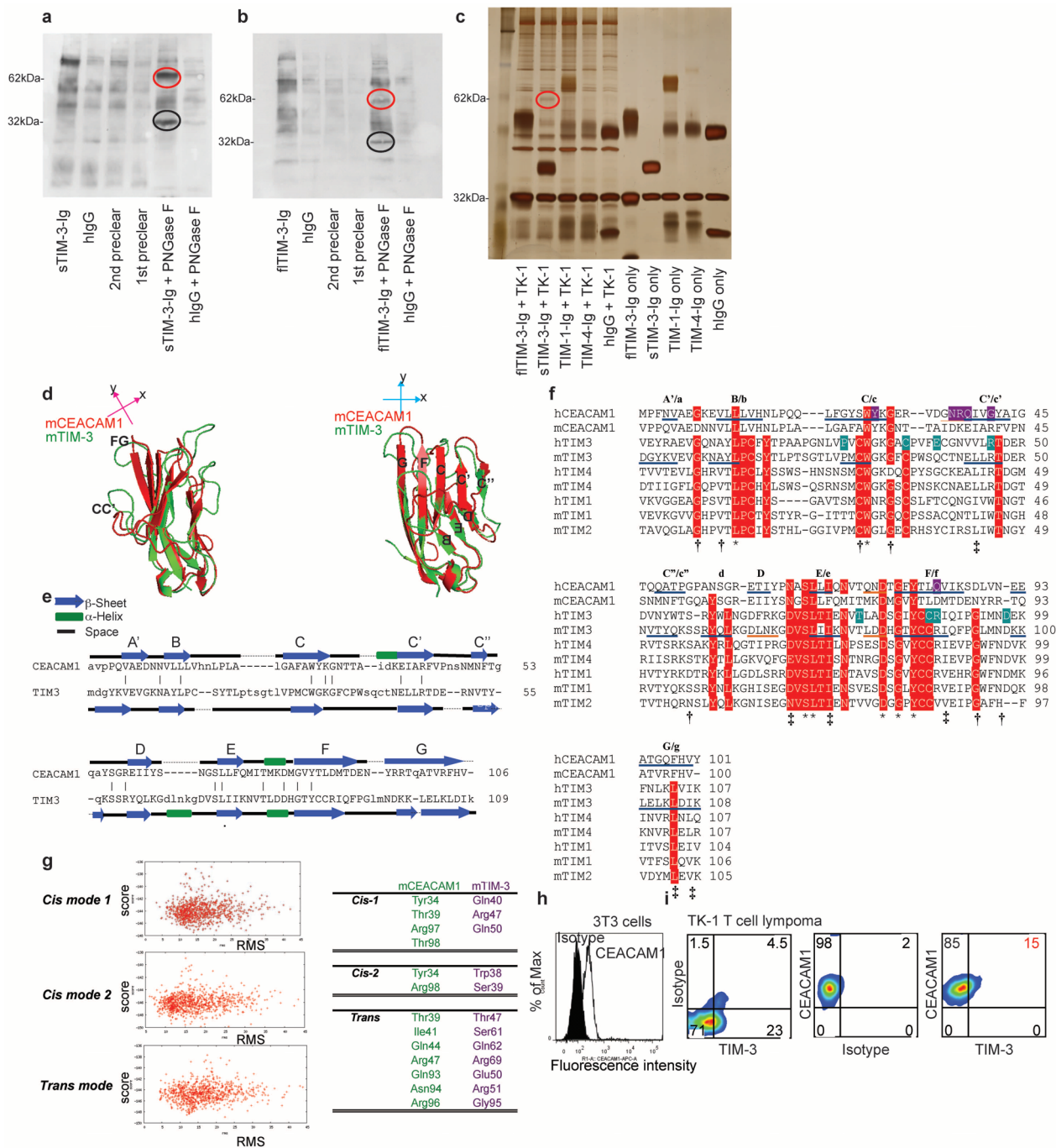




**Extended Data Figure 1 | CEACAM1 is essential for TIM-3 mediated T cell tolerance.** **a**, Schematic diagram of OVA antigen-specific tolerance induction model. **b**, Schematic diagram of OVA immunization. **c**, Tracking *in vivo* antigen-specific T-cell responses of CFSE-labelled OT-II transgenic *Rag2*<sup>-/-</sup> T cells in total lymphocyte gate of mesenteric lymph nodes, peripheral lymph node or spleen of wild-type or *Ceacam1*<sup>-/-</sup> recipients after gating on CFSE-positive cells and staining for CEACAM1 in PBS and OVA<sub>323-339</sub> immunized mice. Hyper-responsiveness of OT-II transgenic *Rag2*<sup>-/-</sup> T cells in *Ceacam1*<sup>-/-</sup> mice was not due to decreased regulatory T-cell induction (data not shown) or increased initial parking on the basis of cell numbers shown. **d**, TIM-3 expression on CEACAM1-positive and -negative CFSE<sup>+</sup> cells as in **c**. **e**, Schematic diagram of SEB-induced T-cell tolerance model. **f**, mCEACAM1 and mTIM-3 expression on CD4<sup>+</sup> Vβ8<sup>+</sup> T cells after SEB tolerance induction. **g**, hCEACAM1 and hTIM-3 expression on activated primary human T cells defined by staining with indicated antibodies. **h**, CEACAM1 expression on TIM-3-silenced primary human T cells after re-activation by flow cytometry. Relative TIM-3, CEACAM1 or CD4 expression on T cells expressing control shRNA (*lacZ* control, red) or three independent shRNAs directed at *TIM3* (overlay, blue). shRNA target sequences shown. **i-l**, CEACAM1 and TIM-3 expression and functional consequences on T cells in HIV infection. CD4<sup>+</sup> IFN-γ<sup>+</sup> T cells are decreased among CEACAM1<sup>+</sup> TIM-3<sup>+</sup> CD4<sup>+</sup> T cells in HIV infection in response to Gag peptides (**i**). Although proportions of CEACAM1<sup>+</sup> TIM-3<sup>+</sup> CD8<sup>+</sup> T cells are similar in HIV-infected and -uninfected subjects (**j**), CEACAM1<sup>+</sup> TIM-3<sup>+</sup> CD8<sup>+</sup> T cells express little

IFN-γ after stimulation with HIV Gag peptides or SEB relative to TIM-3<sup>+</sup> CEACAM1<sup>-</sup> CD8<sup>+</sup> T cells (**k, l**). **C**, hCEACAM1; T, hTIM-3 (*n* = 4 per group, mean ± s.e.m.). **m-o**, *In situ* proximity ligation analysis (PLA) of CEACAM1 and TIM-3. **m**, HEK293T cells transiently co-transfected with Flag-hCEACAM1 or HA-hTIM-3. Cells stained with DAPI (left), anti-tubulin (middle), anti-HA (rabbit) and anti-Flag (mouse) (middle right) or merged (right). Several examples of a positive PLA signal (middle right and right panels: red fluorescent dots) indicative of a maximum distance of 30–40 nm between hCEACAM1 and hTIM-3. **n**, Negative control, co-expression of Flag-PLK1 (protein kinase I) and HA-TIM-3 failed to generate fluorescent dots (that is, PLA negative). Cells stained with DAPI, anti-tubulin, anti-HA/anti-Flag or merged as in **m**. **o**, Negative control, co-expression of HA-ADAP (adhesion and degranulation promoting adaptor protein) failed to show a signal (that is, PLA negative) with staining as in **m**. **p, q**, CEACAM1 and TIM-3 colocalization at immunological synapse of primary human CD4<sup>+</sup> and CD8<sup>+</sup> T cells. Confocal microscopy of hTIM-3<sup>+</sup> hCEACAM1<sup>+</sup> primary CD4<sup>+</sup> and CD8<sup>+</sup> T cells forming conjugates with SEB-loaded B cells. DIC, differential interference contrast. Blue denotes B cell; red denotes CD3; purple denotes CEACAM1; green denotes TIM-3. White indicates colocalization between CEACAM1 and TIM-3 (**p**). Average Pearson correlation coefficients for CD4<sup>+</sup> and CD8<sup>+</sup> T cells were 0.543 and 0.566, respectively, representing strong co-localization (**q**). Data are mean ± s.e.m. and representative of five (**f, g**), four (**p, q**), three (**c, d, m-o**) and two (**h**) independent experiments. \**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.001.

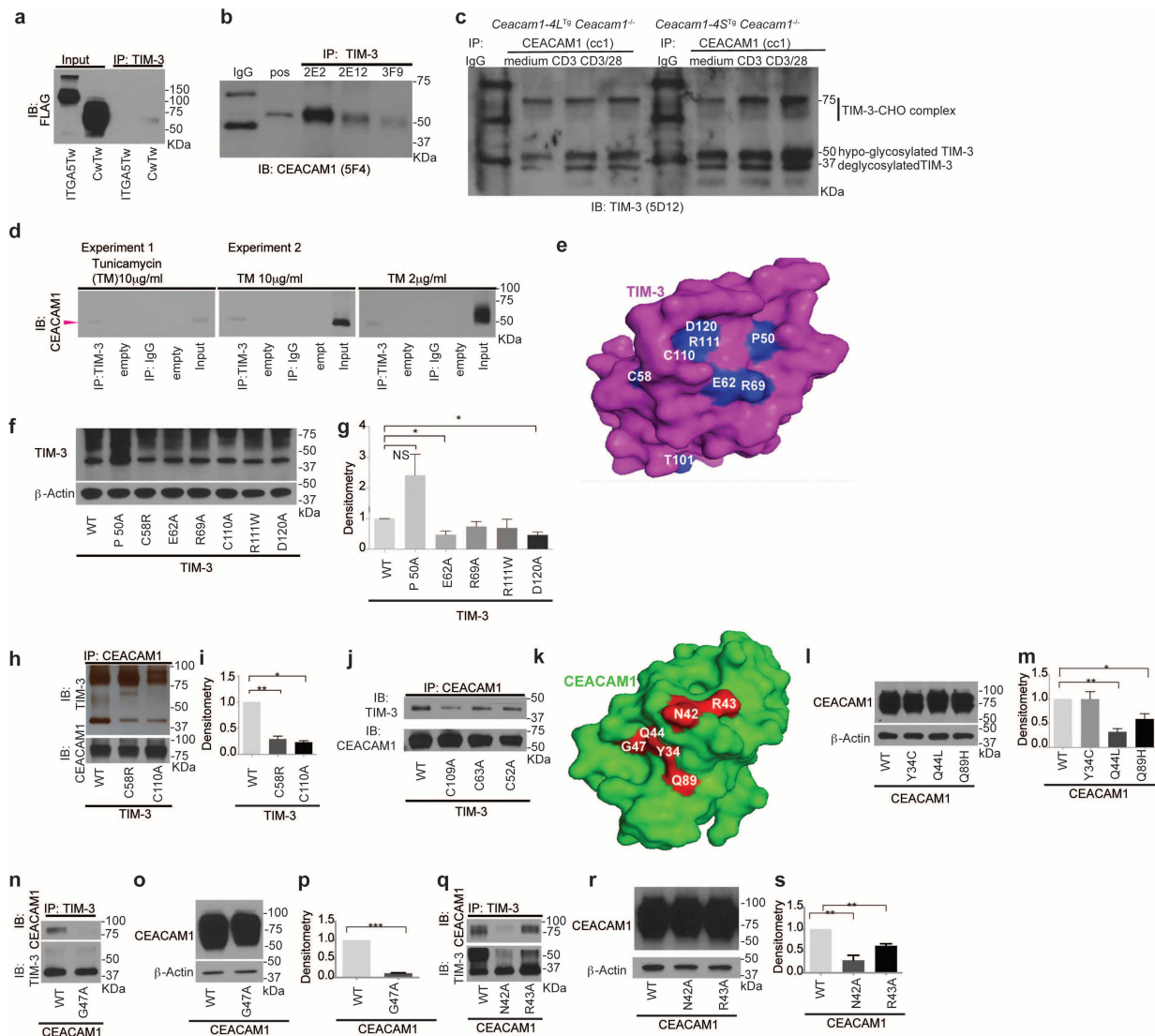




### Extended Data Figure 2 | Structural similarities between CEACAM1 and TIM-3 IgV-like N-terminal domains and biochemical association.

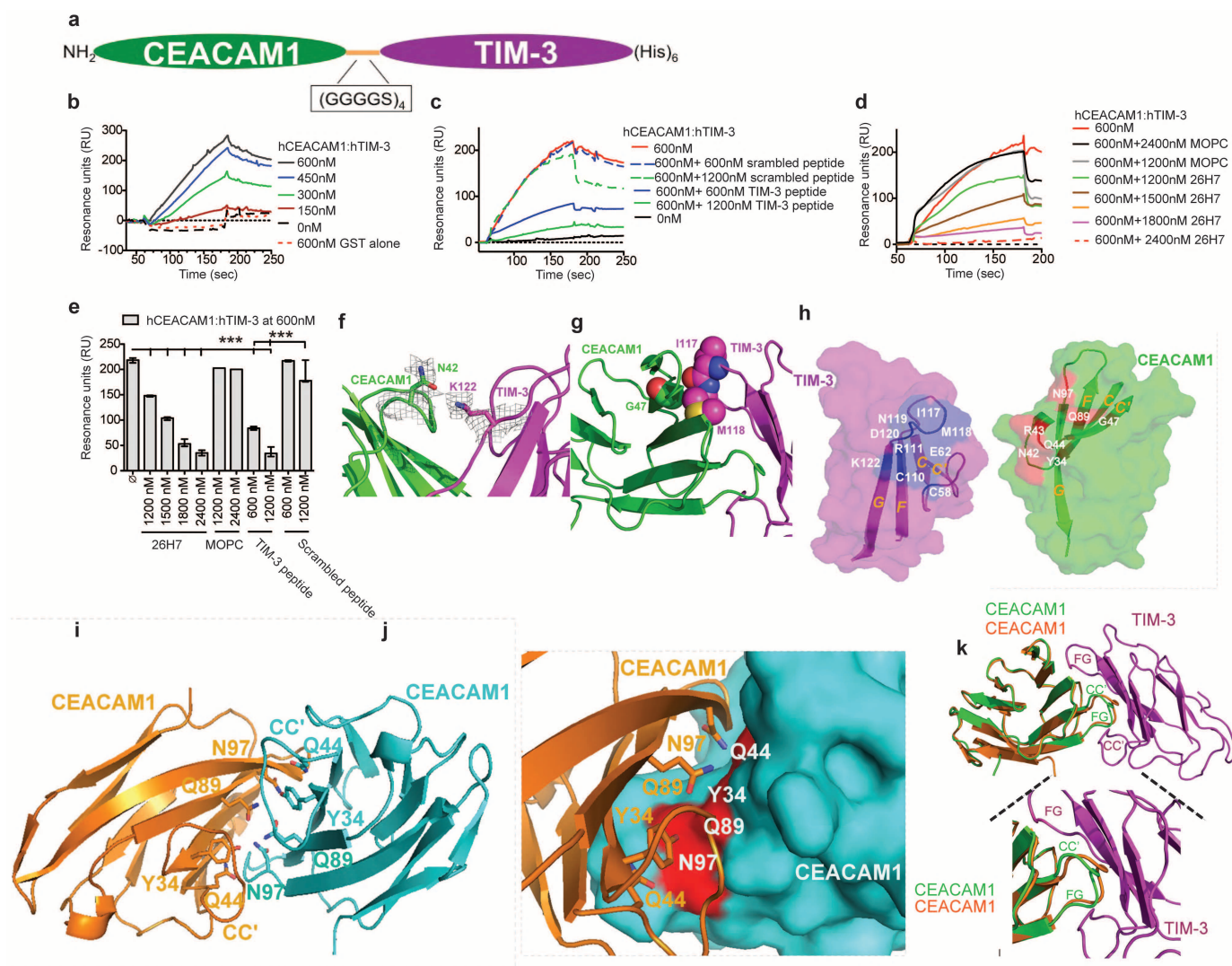
**a–c**, Interaction between TIM-3-Ig fusion protein and membrane protein of 60 kDa after deglycosylation derived from surface-biotinylated TK-1 cells. TIM-3-Ig fusion proteins and human IgG-precipitated proteins were deglycosylated by PNGase F and separated by SDS-PAGE. TIM-3-Ig-binding membrane proteins detected by immunoblot. A 60-kDa membrane protein (red circles) and 32-kDa protein consistent with galectin-9 (black circles) are found specifically associated with soluble (s) TIM-3-Ig fusion protein (**a**, lane 5) and full-length (f) TIM-3-Ig proteins (**b**, lane 5), but not with the pre-clear controls (lanes 3 and 4) or human IgG (lanes 2 and 6). **c**, sTIM-3-Ig and full-length (f) TIM-3-Ig interacting proteins were de-glycosylated by PNGase F and separated by SDS-PAGE. Proteins detected by silver staining. A band of 60 kDa (red circle) is specifically associated with sTIM-3-Ig proteins (lane 2), but not with human IgG (lane 5), or TIM-1-Ig or TIM-4-Ig (lanes 3, 4, 6–10). **d**, Superimposition of previously described IgV-like domains of mCEACAM1 and mTIM-3 demonstrate structural similarity with a score of 2.42 by the structural alignment and root mean square deviation (r.m.s.d.) calculated by

Pymol. **e**, Sequence alignment of the IgV-like domains of mCEACAM1 and mTIM-3 on the basis of the secondary structure alignment in **d**. **f**, Sequence alignments of IgV domain sequences of CEACAM1 and overall mTIM and hTIM family members.  $\alpha$  helices (orange) and  $\beta$  strands (blue) denoted as underlined segments in hCEACAM1 and mTIM-3.  $\beta$  strands labelled with upper- and lower-case letters for hCEACAM1 and mTIM-3, respectively. Conserved residues are shaded red. Mutated residues are shaded violet for hCEACAM1, and green for hTIM-3. Asterisk (\*) indicates positions having a single, fully conserved residue; a dagger (†) indicates conservation between groups of weakly similar residues; a double-dagger (‡) indicates conservation between groups of strongly similar residues. **g**, Computational modelling as defined by energy calculations (score) relative to r.m.s. values of docking models to define potential *cis* and *trans* interfaces between mCEACAM1 and mTIM-3 as described in Supplementary Information and amino acids involved. **h**, CEACAM1 expression on mouse fibroblast 3T3 cells used to identify a galectin-9-independent ligand. **i**, CEACAM1 expression on mouse TK-1 cells as in **a–c**. Representative of three (**a–c**, **h**, **i**) independent experiments.



**Extended Data Figure 3 | Biochemical characterization of interactions between CEACAM1 and TIM-3.** **a**, hTIM-3 does not co-immunoprecipitate (co-IP) with ITGA5 despite interactions with hCEACAM1. HEK293T cells transfected with Flag-ITGA5 and HA-TIM-3 (ITGA5Tw) or Flag-CEACAM1 and HA-TIM-3 (CwTw). Immunoprecipitation with anti-HA antibody and immunoblotting (IB) with anti-Flag antibody are shown. Input represents anti-Flag immunoblot of lysates. **b**, Co-immunoprecipitation of human TIM-3 and CEACAM1 from activated primary human T cells after *N*-glycanase treatment of lysates followed by immunoprecipitation with anti-human TIM-3 antibodies (2E2, 2E12 or 3F9) or IgG as control and immunoblotted with anti-human CEACAM1 antibody (5F4). Protein lysates from HeLa-CEACAM1 transfectants treated with *N*-glycanase followed by immunoprecipitation with 5F4 and the immune complex used as positive control (pos). **c**, mTIM-3 interacts with mCEACAM1 in mouse T cells. Splenocytes from *Ceacam1-4S<sup>tg</sup> Ceacam1<sup>-/-</sup>* and *Ceacam1-4L<sup>tg</sup> Ceacam1<sup>-/-</sup>* mice cultured with anti-CD3 ( $1 \mu\text{g ml}^{-1}$ ) or anti-CD28 ( $1 \mu\text{g ml}^{-1}$ ) or medium for 96 h. Cell lysates immunoprecipitated with anti-mCEACAM1 antibody (cc1) or with mIgG and IB with 5D12 (anti-mTIM-3 antibody) are shown. Locations of mTIM-3 protein variants are indicated. CHO, carbohydrate. **d**, Immunoprecipitation and immunoblot as in **a** with tunicamycin treated, wild-type HA-hTIM-3 and Flag-hCEACAM1 co-transfected HEK293T cells. Arrowhead denotes core CEACAM1 protein. **e**, Potential hCEACAM1-interacting residues on hTIM-3 highlighted in blue. **f**, HEK293 T cells transiently co-transfected with Flag-hCEACAM1 and HA-hTIM-3 mutants. Immunoblotting of anti-HA were used to analyse hTIM-3 expression in HEK293T transfectants. Except for Pro50Ala mutation displaying enhanced overall protein expression, all other mutations in the IgV domain of hTIM-3 are equally detected by anti-HA antibody.

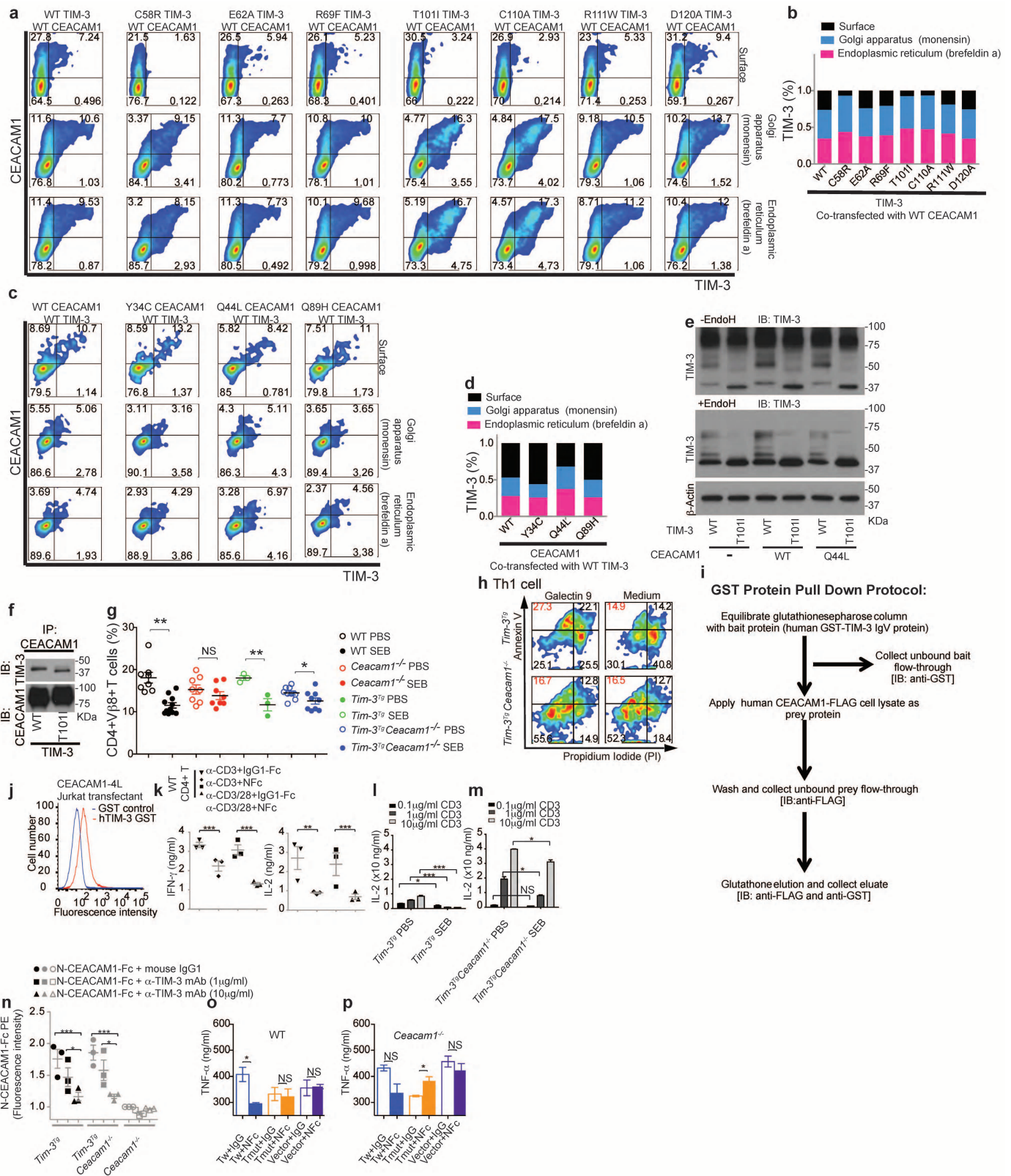
**g**, Quantification of association of hTIM-3 mutants associated with wild-type hCEACAM1 shown in Fig. 2c summing all experiments performed. Association between wild-type hCEACAM1 and hTIM-3 core protein are depicted as reference (set as 1,  $n = 3$ , mean  $\pm$  s.e.m. shown, unpaired Student's *t*-test). **h**, Immunoprecipitation with anti-Flag (hCEACAM1) and immunoblot with anti-HA (hTIM-3) or anti-Flag of wild-type hCEACAM1 and mutant hTIM-3 proteins are shown. **i**, Quantification of **h** as performed in **g**. **j**, HEK293T cells co-transfected with Flag-hCEACAM1 wild-type and HA-hTIM-3 mutants and immunoprecipitation/immunoblot as in **h** revealing no effects of Cys52Ala or Cys63Ala mutations in hTIM-3 in affecting association with hCEACAM1 in contrast to Cys109Ala mutation of hTIM-3 that disrupts interactions with hCEACAM1. **k**, Potential hTIM-3-interacting-residues around the FG-CC' cleft of hCEACAM1 highlighted in red. **l**, HEK293T cells transiently co-transfected with Flag-hCEACAM1 mutants and wild-type HA-hTIM-3. Immunoblot with anti-Flag antibody was used to analyse hCEACAM1 expression in HEK293T co-transfectants. All hCEACAM1 mutations in IgV domain equally detected. **m**, Densitometric quantification of IgV domain hCEACAM1 mutations associating with wild-type HA-hTIM-3 described in Fig. 2d. **n-p**, Analysis of Gly47Ala mutation of hCEACAM1 in hTIM-3 co-transfected HEK293T cells by immunoprecipitation with anti-HA (hTIM-3) and immunoblot with anti-Flag (hCEACAM1) to detect association (**n**), IB with anti-Flag to confirm similarity of hCEACAM1 transfection (**o**) and quantification of associated hCEACAM1 of **n** as shown in **m**. **q-s**, Analysis of hCEACAM1 mutants Asn42Ala and Arg43Ala association with hTIM-3 (**q**), similarity of transfections (**r**) and quantification of **q** as in **n-p**. Representative of four (**d**, **h**), three (**f**, **g**, **i**, **l-s**), two (**a-c**) and one (**j**) independent experiments. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < .001$ .



**Extended Data Figure 4 | Structural analysis of hCEACAM1 and hTIM-3 protein interactions.** **a**, Schematic diagram of single-chain construct consisting of hCEACAM1 IgV-domain (amino acids 1–107), a linker consisting of (GGGGS)<sub>4</sub> and hTIM-3 IgV-domain (amino acids 1–105) and C-terminal hexahistidine tag. **b–e**, Surface plasmon resonance analyses of hCEACAM1–hTIM-3 single-chain interaction with GST–hTIM-3. **b**, Representative sensorgrams of serial dilutions of hCEACAM1–hTIM-3 single chain flowed over immobilized GST–hTIM-3 or GST alone. **c**, Representative sensorgrams of 600 nM hCEACAM1–hTIM-3 single-chain flowed over immobilized GST–hTIM-3 in presence of various concentrations of blocking hTIM-3 specific peptide (amino acids 58–77) or control scrambled peptide. **d**, Representative sensorgrams as in **b** in presence of various concentrations of anti-hCEACAM1 monoclonal antibody (26H7) or control isotype antibody (mIgG1, MOPC). **e**, Bar graphs represent resonance units upon equilibrium (RU<sub>Eq</sub>) of above treatments with mean ± s.e.m. shown from >three runs. GST–hTIM-3 immobilized by amine coupling. Dilutions of

hCEACAM1–hTIM-3 single chain, hCEACAM1–hTIM-3 single chain with either blocking hTIM-3-specific peptide, control scrambled peptide, and 26H7 antibody or control MOPC antibody were injected over immobilized GST–hTIM-3 at 25 °C. Flow rate was 25 µl min<sup>−1</sup>. **f**, **g**, 2F<sub>o</sub> − F<sub>c</sub> maps contoured at 0.9σ showing electron densities for X-ray crystal structure of single chain hCEACAM1–hTIM-3 (PDB code 4QYC). **h**, Summary of crucial amino acid residues defined biochemically and structurally. **i–k**, Similarity between apo-hCEACAM1 and hTIM-3-associated CEACAM1. Structure of CEACAM1 homodimer at 2.0 Å resolution (PDB code 4QXW) (**i**). Homophilic ‘YQQN’ concavity indicated consisting of residues Tyr 34, Gln 44, Gln 89 and Asn 97 at hCEACAM1 (IgV)–hCEACAM1 (IgV) interface (**j**). Superimposition of IgV domain of hCEACAM1 monomer (orange) from **i** on hCEACAM1 (green) from hCEACAM1–hTIM-3 heterodimer in Fig. 2e (**k**). Representative of three (**b–e**) independent experiments. \*\*\*P < 0.001.

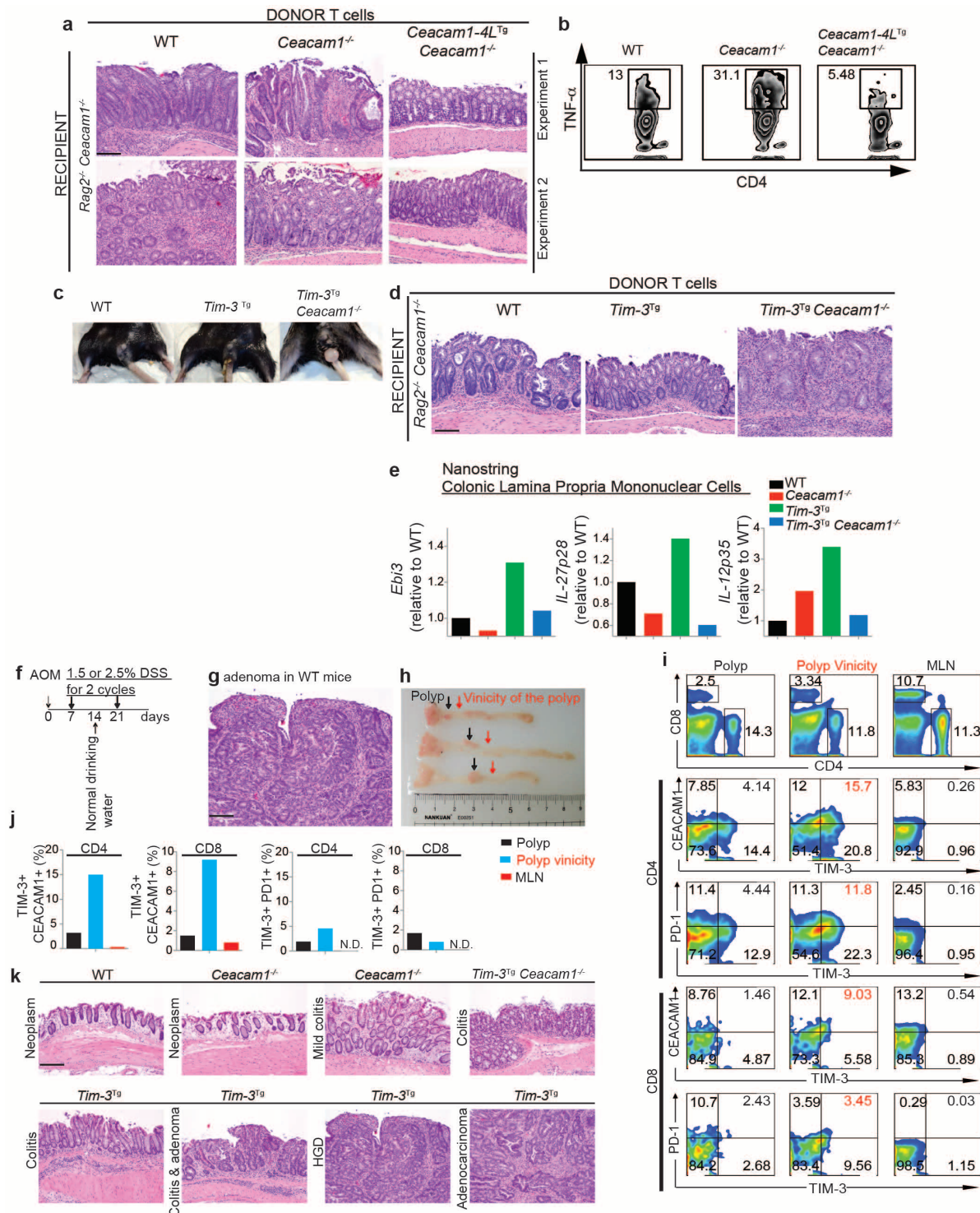






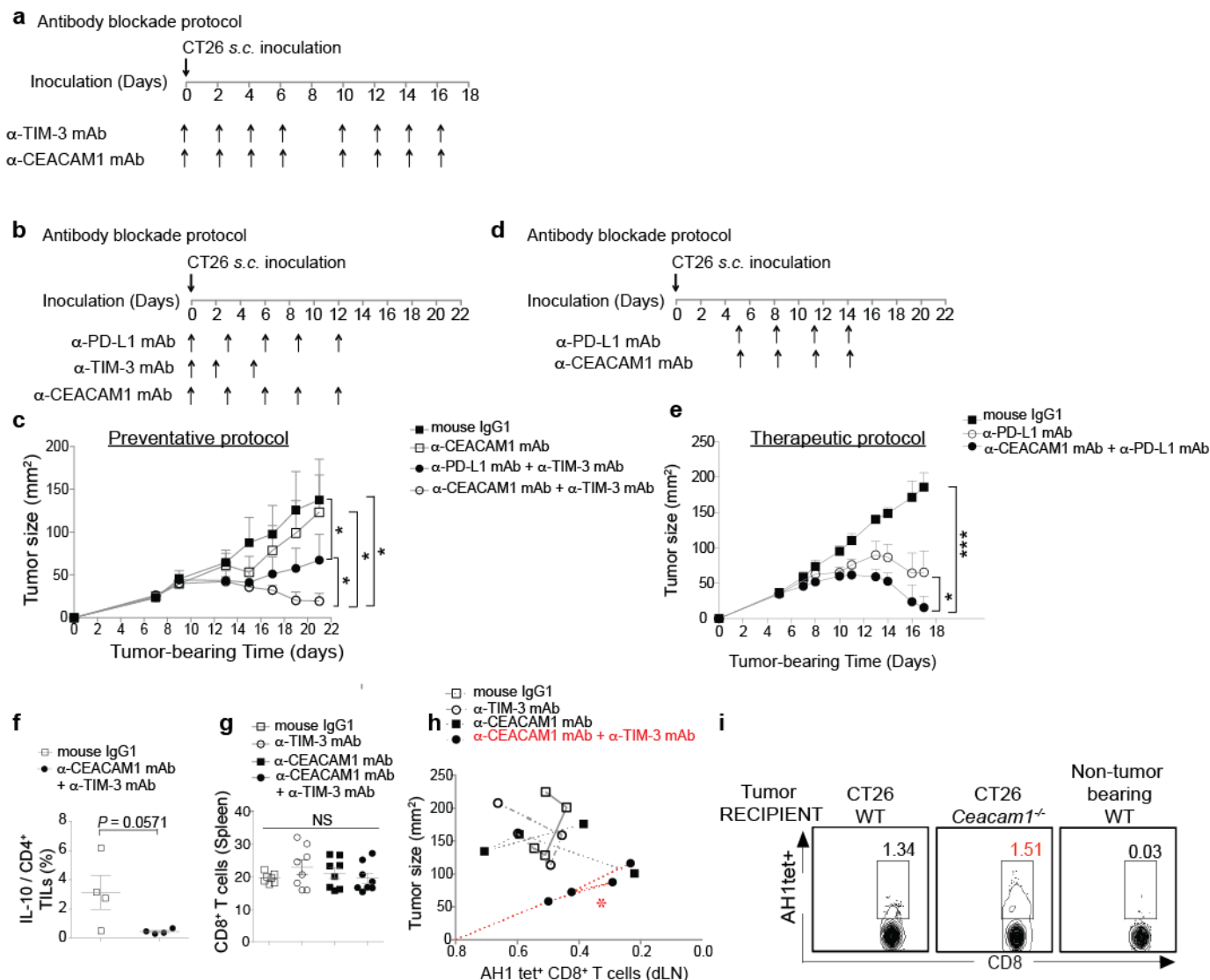
**Extended Data Figure 5 | CEACAM1 determines TIM-3 expression and function.** **a**, HEK293T cells transiently co-transfected with Flag-hCEACAM1 and wild-type or mutants of HA-hTIM-3. Flow cytometry detecting HA-hTIM-3 (detected with anti-HA) and Flag-hCEACAM1 (detected with 5F4) proteins at cell surface (top), Golgi apparatus (middle) or endoplasmic reticulum (bottom) using monensin and brefeldin A, respectively. **b**, Cellular distribution of wild-type or mutant hTIM-3 when co-expressed with wild-type hCEACAM1. Total counts of hTIM-3 at surface, Golgi apparatus and endoplasmic reticulum summed up to 100%. Depicted as percentage of hTIM-3. **c**, HEK293T cells transiently co-transfected with wild-type HA-hTIM-3 (detected with 2E2) and wild-type or mutant Flag-hCEACAM1 (detected with anti-Flag). Flow cytometry analyses as in **a**. **d**, Cellular distribution of **c**, as in **b**. Depicted as percentage of hTIM-3. **e**, Immunoblot for wild-type or Thr101Ile variant of hTIM-3 showing maturation status in presence of wild-type or mutated (Gln44Leu) hCEACAM1. **f**, Normal association of Thr101Ile variant of hTIM-3 with hCEACAM1. **g**, Analysis of CD4<sup>+</sup> V $\beta$ 8<sup>+</sup> T cells after SEB tolerance induction from experimental mice of indicated genotypes. **h**, Galectin-9 induction of apoptosis. Annexin V<sup>+</sup> propidium iodide staining of T<sub>H</sub>1 cells polarized from *Tim3<sup>Tg</sup>* or *Tim3<sup>Tg</sup> Ceacam1<sup>-/-</sup>* mice after treatment with galectin-9 (2  $\mu$ g ml<sup>-1</sup>) for 8 h. Note decreased apoptosis in *Tim3<sup>Tg</sup> Ceacam1<sup>-/-</sup>* T cells. **i**, Schematic diagram of protocol used for protein pull-down using in-column IgV domain of

GST-hTIM-3 incubated with hCEACAM1 protein derived from transfected HEK293T cells as in Fig. 2m. **j**, GST or GST-hTIM-3 staining of hCEACAM1-4L-transfected Jurkat T cells. **k**, Wild-type CD4<sup>+</sup> T cells stimulated with anti-CD3 and/or anti-CD28 in the presence or absence of mCEACAM1 NFc, or IgG1-Fc as control, and cells analysed for secretion of IFN- $\gamma$  and IL-2. **l**, **m**, Characterization of tolerance in SEB model. *Tim3<sup>Tg</sup>* (**l**) and *Tim3<sup>Tg</sup> Ceacam1<sup>-/-</sup>* (**m**) mice treated with SEB with schedule described in Extended Data Fig. 1e. Lymph node cells collected after SEB treatment and re-stimulated with soluble anti-CD3 at indicated doses and IL-2 measured by ELISA after 72 h. Note tolerance in *Tim3<sup>Tg</sup>* but not *Tim3<sup>Tg</sup> Ceacam1<sup>-/-</sup>* mice. *n* = 3 per group. **n**, Anti-mTIM-3 blockade with 2C12 antibody of mCEACAM1 NFc or control IgG-Fc staining of CD4<sup>+</sup> T cells from indicated genotypes expressed as levels relative to *Ceacam1<sup>-/-</sup>* mice. **o**, **p**, Analysis of mTIM-3 cytoplasmic tail function in transmitting mCEACAM1-induced signals. Activated mouse CD4<sup>+</sup> T cells from wild-type (**o**) or *Ceacam1<sup>-/-</sup>* (**p**) mice were retrovirally transduced, sorted and stimulated with anti-CD3 with either human IgG-Fc (IgG, control) or mCEACAM1 N-terminal domain as NFc and TNF- $\alpha$  secretion assessed by ELISA after 72 h. Note ability of CEACAM1 N-terminal domain to transduce a signal associated with inhibition of TNF- $\alpha$  secretion in wild-type but not *Ceacam1<sup>-/-</sup>* T cells. *n* = 3 per group. Data are mean  $\pm$  s.e.m. and represent three (**f**, **g**, **k-p**) and two (**a-e**, **h**, **j**) independent experiments. \**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.001.



**Extended Data Figure 6 | CEACAM1 and TIM-3 cooperatively regulate inflammation and anti-tumour immunity.** **a**, Representative haematoxylin and eosin staining of groups described in Fig. 3e. Scale bar, 50 μm. **b**, Flow cytometry for intracellular cytokine assessment of TNF-α expression from infiltrating CD4<sup>+</sup> T cells from inflamed colonic lamina propria of *Ceacam1*<sup>-/-</sup> *Rag2*<sup>-/-</sup> recipients, 6 weeks after transfer with naive CD4<sup>+</sup> CD44<sup>lo</sup> CD62L<sup>high</sup> T cells from indicated genotypes. **c**, Anorectal prolapse of indicated genotypes. **d**, Representative haematoxylin and eosin staining of groups described in Fig. 3g. Scale bar, 50 μm. **e**, RNA expression defined by nanostring of lamina propria mononuclear cells in indicated groups (mean of *n* = 3 per group). **f**, Schematic overview of protocol for AOM/DSS colitis-associated cancer model. **g**, Representative haematoxylin and eosin staining of colon from wild-type mice in AOM/1.5% DSS model. Scale bar, 50 μm; **h**, Representative photograph of distal colons of wild-type mice (*n* = 3 per

group, anorectal junction at left end) in AOM/1.5% DSS model. Vertical arrows show the sites for dissection of the polyps (black) and the vicinity of the polyps (red). **i**, Representative flow cytometry analyses on infiltrating lymphocytes of invading distal colonic polyps or from the vicinity of the polyps or from mesenteric lymph nodes for CD4<sup>+</sup> and CD8<sup>+</sup> T cells and expression of CEACAM1 and TIM-3 or PD-1 and TIM-3. Note that vicinity of polyps exhibit highest numbers of T cells with an exhausted phenotype. **j**, Summary of flow cytometry on infiltrating lymphocytes from invading distal colonic polyps or from vicinity of polyps and from mesenteric lymph nodes for CD4<sup>+</sup> and CD8<sup>+</sup> T cells expressing CEACAM1 and TIM-3 or PD-1 and TIM-3 (*n* = 3, median shown). **k**, Representative pathology in AOM/1.5% DSS model. Scale bar, 60 μm. HGD, high grade dysplasia. Representative of three independent experiments (a–e, g–k).



**Extended Data Figure 7 | Blockade of CEACAM1 and TIM-3 or genetic loss of CEACAM1 increases anti-tumour immunity.** **a**, Schematic presentation of antibody blockade protocol described in Fig. 4g. **b**, Schematic presentation of antibody blockade protocol referred to in panel c. **c**, Prevention of CT26 tumour growth with indicated combinations of antibodies as in (b) ( $n = 5$  per group, post-hoc Dunnett's correction followed by Friedman test). **d**, Schematic of schedule used for therapeutic antibody administration as described in e. **e**, Synergy of CEACAM1 and programmed death-ligand 1 (PD-L1) blockade in a therapeutic protocol as described in d was performed in wild-type BALB/c mice that received a subcutaneous inoculation of CT26 tumour cells. Mean tumour size ( $n = 5$  per group, with linear regression analysis). Note synergistic increase in anti-tumour effect when CEACAM1 and PD-L1 co-blockade was

performed. **f**, TILs were analysed for the relative proportion of CD4<sup>+</sup> T cells that produced IL-10 as in Fig. 4g ( $n = 4$ , unpaired Student's *t*-test with Mann-Whitney *U* correction). **g**, Percentages of CD8<sup>+</sup> T cells from spleen show that antibody treatments have no effects on total CD8<sup>+</sup> T cell numbers ( $n = 7/8$ , unpaired two-tailed *t*-test). **h**, Negative correlation of the numbers of AH1 tet<sup>+</sup> CD8<sup>+</sup> T cells and the size of tumours in the draining lymph nodes from the tumour-bearing mice in Fig. 4g (Pearson's correlation coefficient,  $r = 0.9560$ ,  $P = 0.044$ ). **i**, Representative flow cytometry for tumour-specific (AH1-tetramer, tet<sup>+</sup>) CD8<sup>+</sup> T cells in draining lymph nodes of mice from the indicated genotypes. Data are mean  $\pm$  s.e.m. and represent three (f–i), two (e) and one (c) independent experiments. \* $P < 0.05$ ; \*\*\* $P < 0.001$ .

## Extended Data Table 1 | Primers for site-directed mutagenesis of hTIM-3 and hCEACAM1

Extended Data Table 1a. TIM-3 mutant primers for site-directed mutagenesis

No.	Mutant Name	Forward primer sequence	Reverse primer sequence
1	P50A	gggaacctcgtggccgtgctgagg	cccagcagacggccacgaggtccc
2	C52A	cctcgtgccgtcgctggggcaaaagg	tccttgccccaggcgacgggcacgagg
3	C58R* (rs201750016)	tggggcaaggagccgctcgtgtttgaatg	cattcaaacacaggacgggctcttggccca
4	E62A	gctgtcctgtgtttgcatgtggcaacgtggtg	caccacgttgccacatgcaaacacaggacaggc
5	C63A	gcaccacgttgcagctcaaacacaggacaggct	agcctgtcctgtgttgaagctggcaacgtggtg
6	R69A	gtggcaacgtggtgctcgcgatgatgaaaggatg	catcccttcatcagtcgagcaccacgttgccac
7	T101I* (rs147827860)	gtccctgacatagaaatgtatttagcagacag	ctgtcgtctagaatcacattctctatggtcaggagac
8	C109A	ggatttggatccggcaggcgtagatccactgtctg	cagacagtgaggatctacgctgcccgaatccaaatcc
9	C110A	cagtgaggatctactgcgcccggatccaaatccca	tgggatttggatccggcgagtagatccactg
10	R111W* (rs145478313)	gggatctactgtgctggtgcaaatccag	ctggatttggatccagcagcagtagatccc
11	D120A	tccaaatccaggcatatgaatgctgaaaaatftaacctgaagt	aacttcaggtttaaattttcagcattatgctctgggatttgg

## b. CEACAM1 mutant primers for site-directed mutagenesis

No.	Mutant name	Forward primer sequence	Reverse primer sequence
1	Y34C* (rs147100915)	ttttgctacagctgtgcaaaagggaaagatgg	ccactctttccccttgcaccagctgtagccaaaa
2	N42A	aggggaagagtggaatggcggccgtcaaatgtaggat	atatcctacaattgacggcgccatccactctttcccct
3	R43A	ggaaaagatggatggcaacgtcaaatgtaggatgca	tgcataatcctacaattgagcgttgccatccactcttcc
4	Q44L* (rs200708090)	gagtggaaggcaaccgtctaattgtaggatgcaata	tattgcataatcctacaattgacggttgccatccactc
5	G47A	ggcaaccgtcaaatgttagcatatgcaataggaaactcaa	ttaggttcctattgcatatgctacaattgacggttgcc
6	Q89H* (rs8111468)	caggattctacaccctacatgtcataaagtcagatcttg	caagatctgactttatgacatgtagggtgtagaatcctg

**a.** hTIM-3 primers. **b.** hCEACAM1 primers. Mutations were chosen on the basis of orthologous human residues reported previously<sup>13</sup> in mTIM-3 as involved in non-galectin-9 interactions with an unknown ligand (hTIM-3 residues 1, 4, 6, 10 and 11 in **a**); orthologous human residues in hCEACAM1 (hCEACAM1 residues 1, 4, 5 and 6 in **b**) predicted to be involved in mTIM-3 interactions on the basis of studies described in Extended Data Fig. 2g and Supplementary Information; natural allelic variants of amino acid residues in hTIM-3 (residues 3, 7 and 10 in **a**) and hCEACAM1 (1, 4 and 6 in **b**) described above or others (hTIM-3 residue 7 in **a**), which were extracted from human exomic databases. Amino acid residues predicted to be involved in hTIM-3 (IgV)–hCEACAM1 (IgV) interactions on the basis of X-ray crystallographic structural models (hCEACAM1 residues 2 and 3 in **b**) as described in Supplementary Information, and cysteine residues involved in intrachain disulphide bonds adjacent to the CC' and FG loops of hTIM-3 (Cys 58–Cys 109, Cys 52–Cys 63 and Cys 38–Cys 110) as described in Supplementary Information. Asterisks represent annotated human natural single nucleotide polymorphisms.



Extended Data Table 2 | Crystal information, data collection and refinement parameters

	hCEACAM1:hTIM-3 single chain	CEACAM1 (Ig-V)
<b>Data collection statistics</b>		
Space Group	P4 <sub>1</sub> 2 <sub>1</sub> 2	P4 <sub>2</sub> 1 <sub>2</sub>
Cell Dimensions (Å)	75.76, 75.76, 147.40	107.16, 107.16, 61.60
Resolution (Å)*	36.85-3.40 (3.72-3.40)	47.92-2.04 (2.11-2.03)
No. of measurements	65773	219575
Unique reflections	6397	23421
I/sigma I	10.2 (6.5)	8.8 (2.3)
Completeness (%)	99.9(100)	99.9 (100)
Redundancy	10.3 (10.1)	9.4 (9.0)
<i>R</i> <sub>merge</sub> (%)	24.5 (49.9)	18.9 (98.4)
<b>Structure refinement</b>		
<i>R</i> <sub>work</sub> (%)	34.5	20.3
<i>R</i> <sub>free</sub> (%)	37.3	24.2
R.m.s deviations		
Bond lengths (Å)	0.014	0.014
Bond angles (°)	1.99	1.23
No. atoms		
Protein	3127	1680
Ligand		47
Water	19	66
B Factors		
Protein	21.20	23.30
Ligand		38.40
Water	30.0	28.60
PDB ID	4QYC	4QXW

\*Highest resolution shell is shown in parenthesis.

**Extended Data Table 3 | Genotype analysis for human *TIM3* alleles in inflammatory bowel disease****a. Study numbers**

Source	Number of individuals	Genotype distribution
German IBD patients	5,598	AA=3/AG=77/GG=5,518
German CD patients	3,975	AA=2/AG=54/GG=3,919
German UC patients	1,623	AA=1/AG=23/GG=1,599
German healthy controls	3,928	AA=0/AG=50/GG=3,878
EVS, European Americans	4,300	AA=0/AG=36/GG=4,264

**c. Clinical data**

Clinical information	Patient		
	1	2	3
Diagnosis	CD	CD	UC
Sex	male	female	male
Age of onset	33	13	40
Disease localisation	Colon	Ileum, colon, rectum	Colon
Disease characteristics	Enterocutaneous fistula and colonic stenosis	Snail-track ulcers, anal stenosis, granulomas, perianal abscesses and fistulas	Perianal fistula
Extraintestinal manifestations	Pancreatitis	Eye, joints, skin	Skin
Treatment	Azathioprine, ASA, steroids	Azathioprine, corticosteroids	unknown
Surgery	Colostomy at age 44, two large bowel surgeries in the following 3 years	Surgery at ages 19 and 21	unknown
Other	Iron and selenium deficiency, hyperuricemia, depression	Chronic active CD, refractory to immunosuppression, hypertension	Family history of IBD

**b. Studied *TIM3* SNPs**

rs-number	CD patients			UC patients			IBD patients			Healthy controls		
	AA	AB	BB	AA	AB	BB	AA	AB	BB	AA	AB	BB
rs201054625	2329	0	0	1448	0	0	3777	0	0	2913	0	0
rs145478313	2302	2	0	1445	0	0	3747	2	0	2875	3	0
rs190484372	2318	0	0	1439	0	0	3757	0	0	2894	0	0
rs147827860	2284	42	2	1423	22	1	3707	64	3	2879	32	0
rs35960726	2305	24	0	1425	22	0	3730	46	0	2861	51	0
rs181855375	2308	0	0	1442	0	0	3750	0	0	2896	0	0
rs184868814	2328	1	0	1447	0	0	3775	1	0	2912	1	0
rs201750016	2275	0	0	1433	0	0	3708	0	0	2839	0	0
rs41283181	2314	1	0	1444	0	0	3758	1	0	2894	0	0
rs147605860	2329	0	0	1447	0	0	3776	0	0	2913	0	0
rs142180056	2329	0	0	1448	0	0	3777	0	0	2913	0	0

Genotype distribution for 11 *TIM3* variants in CD, UC and IBD patients and healthy controls. A specifies the major, B the minor allele.

**a.** Number of successfully genotyped German patients and controls and corresponding distribution of genotypes for the Thr101Ile missense variant in *TIM3*, including the genotype distribution in individuals of European descent (European Americans) downloaded from the Exome Variant Server (EVS), NHLBI GO Exome Sequencing Project (<http://evs.gs.washington.edu/EVS/>). Genotyping 11 rare variants in *TIM3*, in a cohort of 5,634 inflammatory bowel disease and 3,940 control subjects, we observed homozygous rs147827860 (Thr101Ile) carriage exclusively in three IBD subjects (2 Crohn's disease (CD) and 1 ulcerative colitis (UC)), but no control subjects which significantly deviated from Hardy–Weinberg equilibrium in IBD cases ( $P = 0.0033$ ) but not controls ( $P = 1.00$ ). **b.** Description of *TIM3* single nucleotide polymorphisms (SNPs) analysed in cohorts described in **a.** **c.** Clinical data for the three patients found to be homozygous for the rs147827860 (Thr101Ile) variant.

# An ERK/Cdk5 axis controls the diabetogenic actions of PPAR $\gamma$

Alexander S. Banks<sup>1</sup>, Fiona E. McAllister<sup>2</sup>, João Paulo G. Camporez<sup>3</sup>, Peter-James H. Zushin<sup>1</sup>, Michael J. Jurczak<sup>3</sup>, Dina Laznik-Bogoslavski<sup>4</sup>, Gerald I. Shulman<sup>3</sup>, Steven P. Gygi<sup>2</sup> & Bruce M. Spiegelman<sup>2,4</sup>

**Obesity-linked insulin resistance is a major precursor to the development of type 2 diabetes. Previous work has shown that phosphorylation of PPAR $\gamma$  (peroxisome proliferator-activated receptor  $\gamma$ ) at serine 273 by cyclin-dependent kinase 5 (Cdk5) stimulates diabetogenic gene expression in adipose tissues<sup>1</sup>. Inhibition of this modification is a key therapeutic mechanism for anti-diabetic drugs that bind PPAR $\gamma$ , such as the thiazolidinediones and PPAR $\gamma$  partial agonists or non-agonists<sup>2</sup>. For a better understanding of the importance of this obesity-linked PPAR $\gamma$  phosphorylation, we created mice that ablated Cdk5 specifically in adipose tissues. These mice have both a paradoxical increase in PPAR $\gamma$  phosphorylation at serine 273 and worsened insulin resistance. Unbiased proteomic studies show that extracellular signal-regulated kinase (ERK) kinases are activated in these knockout animals. Here we show that ERK directly phosphorylates serine 273 of PPAR $\gamma$  in a robust manner and that Cdk5 suppresses ERKs through direct action on a novel site in MAP kinase/ERK kinase (MEK). Importantly, pharmacological inhibition of MEK and ERK markedly improves insulin resistance in both obese wild-type and *ob/ob* mice, and also completely reverses the deleterious effects of the Cdk5 ablation. These data show that an ERK/Cdk5 axis controls PPAR $\gamma$  function and suggest that MEK/ERK inhibitors may hold promise for the treatment of type 2 diabetes.**

Obesity is characterized by dysfunctional adipose tissues in which failure to store excess energy appropriately leads to ectopic lipid deposition, progressive insulin resistance and heightened risk for type 2 diabetes. Disordered secretion of certain fat-derived hormones, called adipokines, also contributes to the metabolic dysfunction in obesity and diabetes. Adipose-tissue-directed insulin-sensitizing drugs, including the thiazolidinediones, potentially improve whole-body sensitivity to insulin<sup>3</sup>. The thiazolidinedione drugs have two distinct functions as ligands for PPAR $\gamma$ : they promote the differentiation of preadipocytes<sup>4,5</sup> and they block the phosphorylation of PPAR $\gamma$  at S273 (ref. 1). We recently demonstrated that non-agonist PPAR $\gamma$  ligands capable of blocking the phosphorylation of PPAR $\gamma$  at S273 retain potent anti-diabetic effects despite the inability to promote adipogenesis<sup>2</sup>. These findings strongly suggested that obesity-mediated phosphorylation of PPAR $\gamma$  S273 may not only correlate positively with the development of insulin resistance but may also be causal to this state.

A variety of protein kinases participate in insulin action and insulin resistance. Insulin signalling activates the Akt/phosphoinositide 3-kinase (PI(3)K) and Grb2/Ras/MEK/ERK kinase cascades<sup>6,7</sup>. Although much is known about the role of the former in promoting the canonical anabolic actions of insulin, studies *in vitro* had suggested that the latter cascade downstream of insulin signalling could contribute to insulin resistance<sup>8,9</sup>, although there is controversy on this point<sup>10</sup>. Obese rodents were shown to have elevated ERK activity, whereas mice lacking ERK1 were shown to be more sensitive to the effects of insulin<sup>9,11,12</sup>.

Cdk5 function is both necessary and sufficient in cultured adipocytes to phosphorylate PPAR $\gamma$  at S273 (ref. 1). Mice with global or

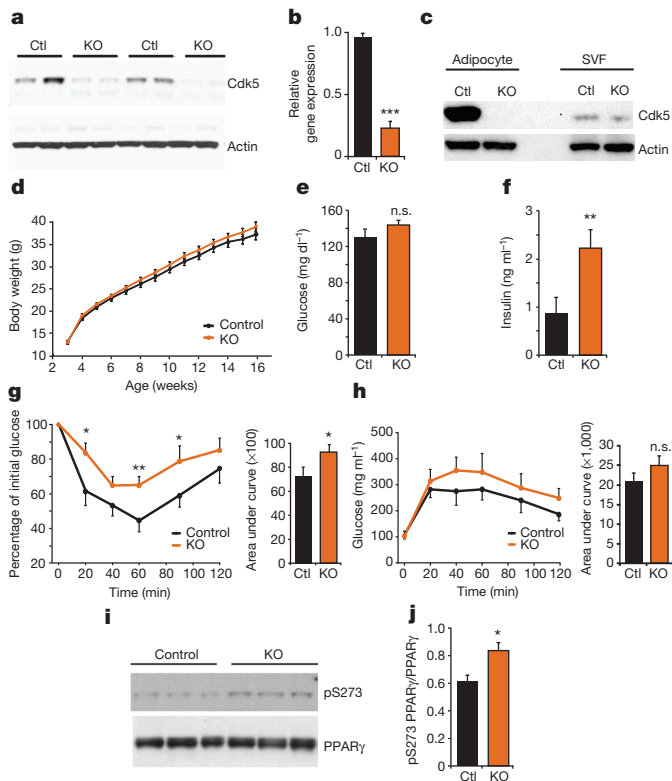
brain-restricted deletion of Cdk5 show increased perinatal mortality due to a defect in neurogenesis. We therefore set out to test whether modulation of PPAR $\gamma$  phosphorylation *in vivo* in adipose tissues would lead to altered insulin sensitivity *in vivo* by creating adipose-selective Cdk5-deficient mice (Cdk5-KO)<sup>13,14</sup>. In contrast to global knockouts<sup>15,16</sup>, Cdk5-KO mice are grossly normal in appearance, with no apparent differences in body weight or fasting glucose levels when maintained on a standard diet (Extended Data Fig. 1). Deletion of Cdk5 in whole white adipose tissue was confirmed by both western blot analysis (Fig. 1a) and quantitative real-time PCR (Fig. 1b). To determine whether the residual Cdk5 expression in the knockout (KO) mice was emanating from non-adipocytes or from incomplete recombination, tissue fractionation was performed; no detectable Cdk5 protein was observed in the floating adipocyte fraction, whereas residual signal was observed in the stromal vascular fraction (Fig. 1c). On a standard chow diet, KO mice were normal, healthy and indistinguishable from Cdk5<sup>Flox/Flox</sup> controls (Extended Data Fig. 1).

Both PPAR $\gamma$  S273 phosphorylation and insulin resistance are strongly promoted by obesity and inflammatory cytokines<sup>1</sup>. When maintained on a high-fat diet to induce obesity, no differences were observed between wild-type (WT) and KO groups in food intake, energy expenditure or body weight (Fig. 1d and Extended Data Fig. 2). Paradoxically, metabolic analyses of these Cdk5-KO mice demonstrated that their glucose homeostasis was impaired in comparison with control animals. Cdk5-KO mice had elevated fasting insulin levels, as well as impairment in insulin tolerance, with a trend towards impaired glucose tolerance (Fig. 1e–h). We also observed a paradoxical increase in S273 PPAR $\gamma$  phosphorylation in obese Cdk5-KO mice, strongly suggesting compensation from an alternative protein kinase (Fig. 1i, j).

To understand how PPAR $\gamma$  S273 phosphorylation is increased in the absence of Cdk5, unbiased quantitative proteomic kinase profiling was performed on white adipose tissue (Fig. 2a). The most enriched protein kinase-derived peptide in KO mice (VADPDHDHTGFLTEpY<sup>185</sup>VATR) corresponded to the activation loop of MAP kinase, ERK2/Mapk1 (Fig. 2b). We independently confirmed that ERK2 was activated in adipose tissue extracts from the KO mice by examining the phosphorylation of ERK2 at T183 and Y185, using phospho-specific antibodies against ERKs (Fig. 2c). We found no significant differences in the activation of alternative obesity-linked kinases (Extended Data Fig. 3). In addition, elevated T183/Y185 ERK phosphorylation was observed in adipocytes from Cdk5-KO mice that were differentiated for 8 days *in vitro*, suggesting that this phenomenon is cell-autonomous (Fig. 2d). A small-molecule inhibitor of Cdk5, roscovitine<sup>17</sup>, also promotes ERK phosphorylation in cultured adipocytes, although here the competing inhibition of Cdk5 and activation of ERK had a net effect of leaving PPAR $\gamma$  S273 phosphorylation unchanged (Fig. 2e).

Because even the best available inhibitors of Cdk5, such as roscovitine, are not completely specific for Cdk5 (refs 17, 18), we also used a more precise means of regulating this kinase: an allele of Cdk5 specifically

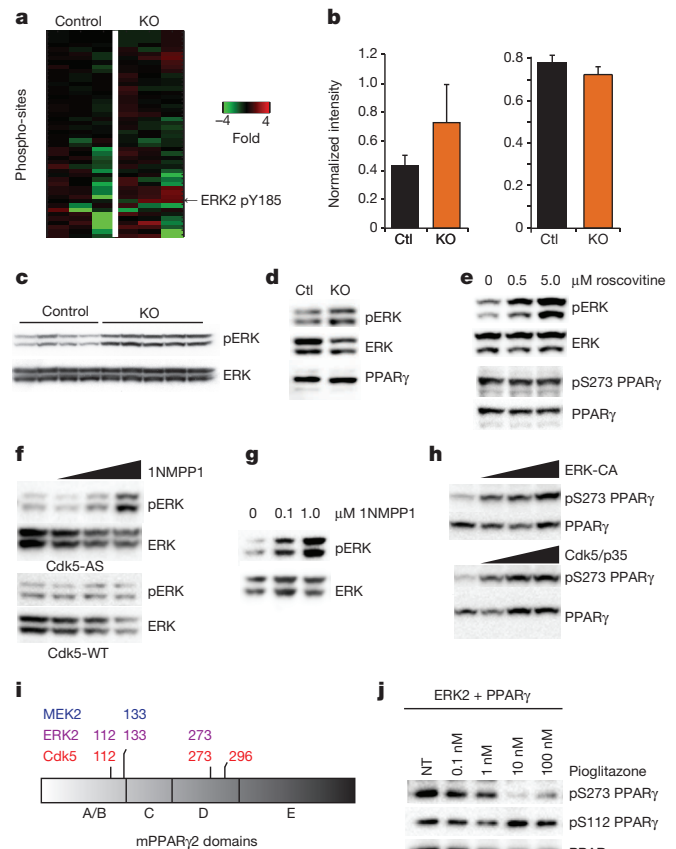
<sup>1</sup>Division of Endocrinology, Diabetes and Hypertension, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>2</sup>Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>3</sup>Yale Mouse Metabolic Phenotyping Center and Departments of Internal Medicine and Cellular and Molecular Physiology, Yale University School of Medicine, New Haven, Connecticut 06510, USA. <sup>4</sup>Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA.



**Figure 1 | Insulin resistance after Cdk5 deletion in adipocytes.** **a, b**, Deletion of Cdk5 in epididymal white adipose tissue from control (Cdk5<sup>Flox/Flox</sup>) or adipocyte-specific knockout, KO (Cdk5<sup>Flox/Flox</sup>::adiponectin-Cre) was confirmed by western blotting (**a**;  $n = 4$ ) or quantitative real-time PCR (**b**;  $n = 5$ ). Ctl, control. **c**, Fractionated adipose tissue confirmed deletion was confined to the adipocyte fraction of adipose tissue. **d**, Body weight of control or KO mice when maintained on a high-fat diet.  $n = 20$  Ctl, 25 KO. **e, f**, Fasting glucose (**e**) and fasting insulin (**f**) in mice maintained on a high-fat diet.  $n = 10$  (control) and 12 (KO). **g, h**, Insulin tolerance test (**g**) and glucose tolerance tests (**h**) are consistent with impaired insulin sensitivity.  $n = 15$  (control) and 17 (KO). Histograms show the areas under the curves. **i**, Western blots of white adipose tissue for pS273 PPARγ in control and KO mice. **j**, Quantification of **i**. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ ; n.s., not significant. Error bars indicate s.e.m.

designed to be inhibited by the 'bulky' small molecule 1NMPP1. With the use of this previously validated approach<sup>19,20</sup>, a dose-dependent increase of ERK phosphorylation was observed after Cdk5 inhibition with 1NMPP1, in both fibroblasts and cultured adipocytes (Fig. 2f, g). Thus, we observed elevated levels of an activating ERK phosphorylation as a consequence of Cdk5 loss *in vivo* or *ex vivo*.

We next investigated whether ERK kinase might be capable of directly compensating for Cdk5 deficiency in the phosphorylation of PPARγ. ERK and Cdk5 are structurally similar Ser/Thr kinases with a propensity for phosphorylating sites with proline in the +1 position<sup>21</sup>. In cultured cells, both constitutively active ERK (ERK-CA) and active Cdk5 (Cdk5 and its activating subunit p35) phosphorylated PPARγ at S273 (Fig. 2h). We confirmed that this was a direct effect by performing *in vitro* protein kinase assays on recombinant full-length PPARγ. Both ERK2 and Cdk5 resulted in the direct phosphorylation of S273 PPARγ, but a third kinase, MEK2, failed to phosphorylate this site (Fig. 2i). Cdk5 and ERK both phosphorylated the Ser-Pro sites S273 and S112 (Fig. 2i). The action of ERK on S112 has been reported previously by us and others<sup>22–24</sup>. A novel Cdk5-specific target site at T296 was also identified. In contrast, both ERK and MEK phosphorylated PPARγ at only one common site, S133. Because thiazolidinediones can block the ability of CDK5 to phosphorylate S273 PPARγ (ref. 1), we sought to determine whether they would similarly block the action of ERKs. ERK phosphorylates both S112 and

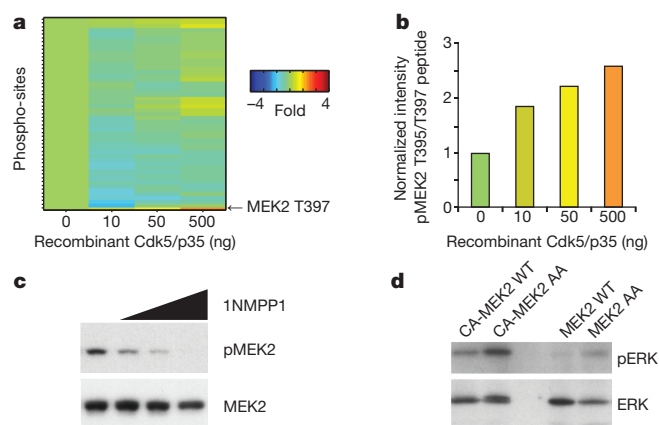


**Figure 2 | Identification and characterization of ERK as a S273 PPARγ kinase.** **a**, ATP-probe-enriched phosphoproteomic analysis of kinases in control or Cdk5-KO adipose tissue from mice fed on a high-fat diet. Heat map indicating the most highly regulated phosphopeptide between control and Cdk5-KO ( $n = 3$  per group;  $P = 0.08$ ) corresponding to the peptides containing the activation loop of Mapk1/ERK2. **b**, Normalized quantification of the abundance of the phosphopeptide (left) and total abundance (right) corresponding to the activation loop of ERK2, determined by mass spectrometry. **c**, Western blot of ERK1 and ERK2 phosphorylation in brown adipose tissue from mice on a high-fat diet. **d**, Western blot of phospho-ERK1/2 in primary adipocytes differentiated *in vitro* for 8 days and serum starved for 18 h. **e**, Inhibition of Cdk5 by treatment with roscovitine for 6 h in cultured F442A adipocytes at the indicated doses. **f**, HEK293 cells expressing WT Cdk5 or an analogue-sensitive (AS) mutant of Cdk5 were treated with the AS-specific inhibitor 1NMPP1 at 0, 0.1, 1.0 or 10 μM for 2 h. **g**, Cultured adipocytes stably expressing Cdk5-AS treated with the indicated doses of 1NMPP1. **h**, PPARγ in HEK293 cells co-transfected with increasing doses of constitutively active ERK2 kinase (ERK-CA) or active Cdk5 (Cdk5 with p35). Western blotting was performed for both pS273 PPARγ and total PPARγ. **i**, Phosphorylated residues identified by LC-MS/MS after *in vitro* kinase assay of recombinant Cdk5, ERK2, or MEK2 incubated with full-length recombinant PPARγ. **j**, *In vitro* ERK kinase assay with incubated with full-length recombinant PPARγ and increasing doses of pioglitazone before western blotting for pS273 and total PPARγ. NT, no treatment. Error bars indicate s.e.m.

S273, but increasing concentrations of pioglitazone block phosphorylation only at S273 (Fig. 2j).

To determine how Cdk5 might be regulating ERK, we again turned to ATP probes and quantitative proteomics to identify Cdk5 substrates. Cdk5-deficient adipose tissue extracts were spiked with increasing doses of recombinant active Cdk5 kinase, plus p35 (Fig. 3a). MEK2, the kinase upstream of ERK, was identified as the protein with the single greatest dose-dependent increase in phosphorylation (Fig. 3b). The phosphopeptide identified contained two closely spaced potential phosphothreonine sites at T395 and T397 of mouse MEK2. Although Cdk5 has been reported to regulate MEK1 at T286, this site was not conserved in MEK2





**Figure 3 | Regulation of MEK2 by Cdk5.** **a**, ATP-probe phosphoproteomic analysis of activated kinases in Cdk5-KO adipose tissue lysates after addition of recombinant active Cdk5 and p35 protein at the indicated doses. **b**, The most highly regulated peptide includes MEK2 residues T395 and T397. **c**, HEK293 cells expressing the AS mutant of Cdk5 were treated with the AS-specific inhibitor 1NMPP1 at 0, 0.1 1.0 or 10  $\mu$ M for 2 h before western blotting for MEK2 phospho-T395. **d**, Immunoprecipitation *in vitro* kinase assay of MEK2-WT or T395A T397 (MEK2-AA) mutant. Flag-tagged WT-MEK2 or constitutively active CA-MEK2 (S222D S226D) were immunoprecipitated from HEK293 cells and incubated with ATP and recombinant ERK protein.

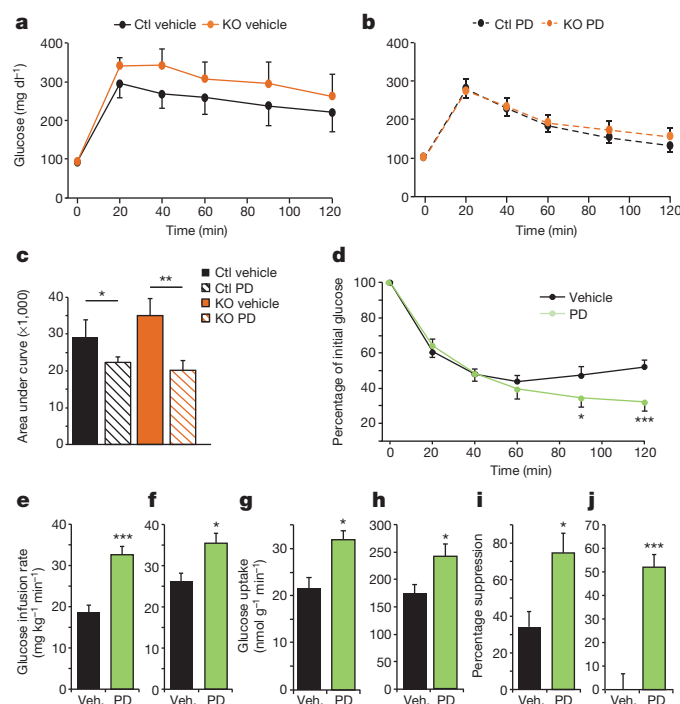
(Extended Data Fig. 4)<sup>25</sup>. The function of these two MEK2 sites (T395 and T397) has not previously been reported, yet they are the most frequently found MEK2 modifications in the proteomic databases outside the canonical activation loop<sup>26</sup>.

Using a phospho-specific antibody, we were able to show that acute Cdk5 inhibition abrogates the phosphorylation of MEK2 at T395 (Fig. 3c). Mutation of both of these neighbouring sites rendered MEK more active than wild type in a protein kinase assay using recombinant ERK protein as a substrate (Fig. 3d). Together, these findings strongly suggest that Cdk5 deletion results in the activation of ERK kinase via derepression of MEK kinase activity.

We next asked a critical question: does inhibition of the MEK/ERK pathway correct the metabolic defect evident in the Cdk5-KO mice? Mice of both WT and KO genotypes maintained on a high-fat diet were treated with the well-characterized MEK inhibitor PD0325901 for 5 days before a glucose tolerance test. Inhibition of MEK was able to normalize glucose tolerance completely in these two groups, consistent with the role of ERK as a key compensating kinase in Cdk5-deficient adipose tissue (Fig. 4a–c). This occurred with no effect on adiposity (Extended Data Fig. 5).

To gain a better understanding of the role of ERK in the pathophysiology of diabetes we examined insulin sensitivity by performing insulin tolerance tests and hyperinsulinaemic–euglycaemic clamp experiments on diet-induced obese wild-type C57Bl/6 mice. Mice treated with PD0325901 showed a late divergence in glycaemia 90 min after administration of insulin, suggesting an exaggerated response to insulin action (Fig. 4d). Similarly, clamp studies revealed a twofold increase in the glucose infusion rate (Fig. 4e). This markedly improved sensitivity to insulin infusion was due to increased whole-body glucose utilization and improved insulin-mediated suppression of both endogenous glucose production and lipolysis (Fig. 4f, i, j). Tracer analysis indicated that insulin sensitivity was increased more strongly in adipose tissue than in skeletal muscle; however, both tissues contributed to the effects of ERK inhibition on glucose disposal (Fig. 4g, h).

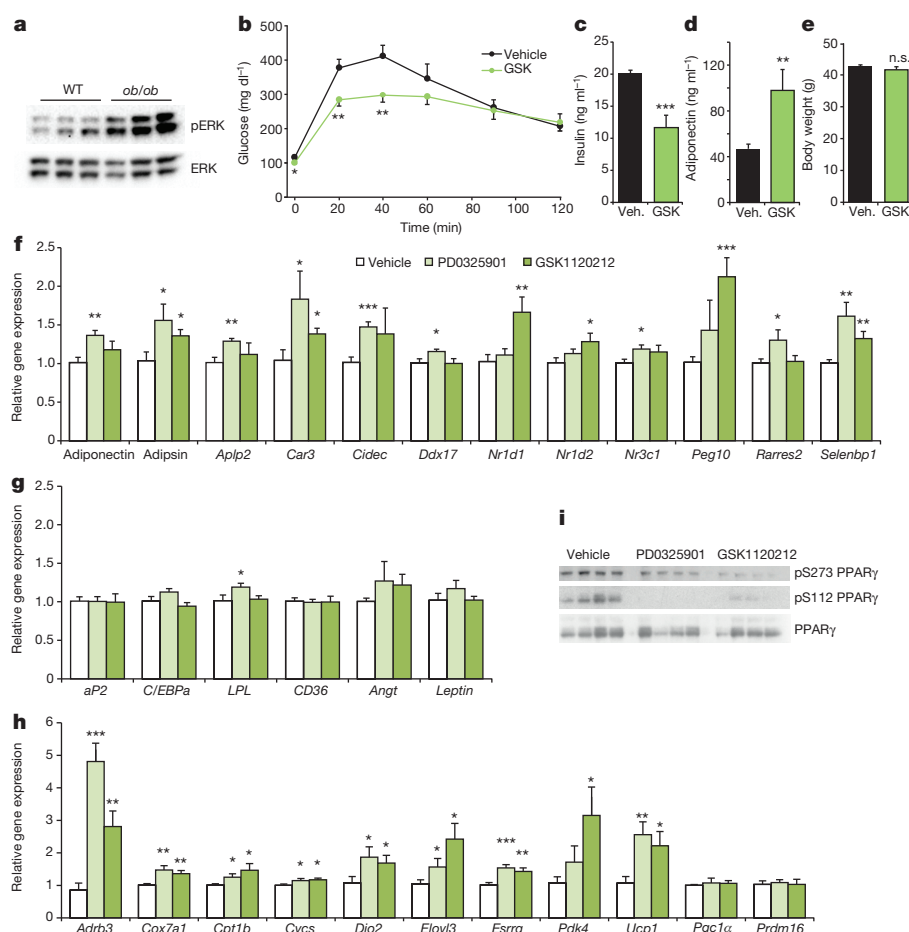
To examine further the efficacy of targeting MEK and ERK pharmacologically, we studied severely insulin-resistant leptin-deficient obese (*ob/ob*) mice. We found elevated phospho-ERK levels in white adipose tissue from unchallenged *ob/ob* mice in comparison with lean WT controls



**Figure 4 | Metabolic consequences of MEK inhibition *in vivo* in mice fed with a high-fat diet.** **a**, **b**, Glucose tolerance tests of high-fat-diet-fed adipose-specific Cdk5-KO mice or controls after treatment with vehicle (**a**) or MEK inhibitor PD0325901 (PD) (**b**). **c**, Area under the curves in **a** and **b**.  $n = 10$  for Ctl vehicle, KO vehicle and KO PD groups;  $n = 9$  for the Ctl PD group. **d**, Insulin tolerance tests ( $n = 12$  Ctl, 11 KO) in high-fat-diet-fed wild-type C57Bl/6 mice after treatment with PD0325901. **e**–**j**, Hyperinsulinaemic–euglycaemic clamps in high-fat-diet-fed wild-type C57Bl/6 mice after treatment with PD0325901. **e**, Glucose infusion rate. Veh., vehicle. **f**, Whole-body glucose uptake/disposal. **g**, <sup>3</sup>H-2-deoxyglucose (2DG) tracer uptake into epididymal white adipose tissue. **h**, 2DG tracer uptake into gastrocnemius. **i**, Percentage suppression of endogenous glucose production. **j**, Percentage suppression of free fatty acids.  $n = 9$  vehicle, 11 PD. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ . Error bars indicate s.e.m.

(Fig. 5a). We next treated *ob/ob* mice with either PD0325901 or a distinct US Food and Drug Administration-approved MEK inhibitor, Trametinib/GSK1120212. In comparison with control animals, mice receiving either GSK1120212 or PD0325901 showed an improvement in glucose tolerance (Fig. 5b and Extended Data Fig. 6a). This was accompanied by decreased insulin levels and increased levels of the insulin-sensitizing hormone adiponectin (Fig. 5c, d and Extended Data Fig. 6b) without affecting body weight (Fig. 5e and Extended Data Fig. 6c).

We performed gene expression analysis to gain a better understanding of the transcriptional basis of the improved glucose homeostasis after MEK inhibition. We have previously defined a set of 17 genes that are sensitive to PPAR $\gamma$  S273 phosphorylation in cultured adipose cells and then further refined that gene set to 10 genes that were also regulated by obesity in mice and responded to treatment with non-agonist PPAR $\gamma$  ligands<sup>1</sup>. Although treatment with MEK inhibitors did not affect the degree of obesity, expression of all of these 10 genes was significantly regulated by treatment with one or both of the MEK inhibitors used in this study (Fig. 5f). This included the genes encoding the circulating insulin sensitivity factors adiponectin and adiponin, previously shown to be most sensitive to phosphorylation of PPAR $\gamma$  at S273. The direction of all of these changes was consistent with that predicted by the reversal of S273 phosphorylation. Conversely, the MEK inhibitors had a minimal effect on the expression of genes linked to PPAR $\gamma$  agonism and adipogenesis, including *aP2* and *Cebpa* (Fig. 5g). We also found increased expression of genes participating in the induction of thermogenesis and ‘browning’ of white adipose tissue (Fig. 5h). This thermogenic program



**Figure 5 | Metabolic consequences of MEK inhibition in *ob/ob* mice.** **a**, ERK phosphorylation in white adipose tissue from WT C57Bl6/J and from *ob/ob* mice. **b**, Glucose tolerance test in *ob/ob* mice after treatment with the MEK inhibitor GSK1120212 (GSK) or vehicle. **c–e**, Fasting insulin values (**c**), plasma adiponectin levels (**d**) and body weights (**e**) ( $n = 8$ ). **f–h**, Gene expression in *ob/ob* epididymal white adipose tissue after treatment with vehicle or either of two MEK inhibitors, PD0325901 or GSK1120212 ( $n = 7, 7$  and  $8$ , respectively). **f**, Genes responsive to PPAR $\gamma$  S273 phosphorylation. **g**, Genes responsive to PPAR $\gamma$  agonism. **h**, Genes controlling 'browning' of white adipose tissue and thermogenesis. **i**, Phosphorylation of PPAR $\gamma$  in epididymal white adipose tissue in *ob/ob* mice after treatment with MEK inhibitors. Areas under the curve and gene expression were analysed by analysis of variance. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ ; n.s., not significant. Error bars indicate s.e.m.

does not include induction of the PPAR $\gamma$  coactivators *Pgc1 $\alpha$*  or *Prdm16*, although post-translational modifications of these proteins cannot be excluded as contributing to this phenotype<sup>27,28</sup>. Last, we observed decreased expression of the pro-inflammatory cytokine *Tnfx* messenger RNA and an altered adipose tissue macrophage expression profile (Extended Data Fig. 7). Both MEK inhibitors caused a decrease in PPAR $\gamma$  phosphorylation at S112 and strongly suggesting a new role in regulating S273 (refs 22–24) (Fig. 5i). Taken together, these data identify an insulin-sensitizing role for MEK inhibitors in adipose tissue, which is consistent with the effects of non-agonist PPAR $\gamma$  ligands that specifically block PPAR $\gamma$  phosphorylation at S273 (Extended Data Fig. 8).

The MEK inhibitory compounds that we have used here are now safe and effective enough to be used in patients with metastatic melanoma<sup>29</sup> and are tolerated well enough to permit studies of metabolism in rodents and perhaps in humans. We find anti-diabetic effects when these MEK inhibitors are given at doses threefold lower than were used in rodent tumour xenograft models<sup>30</sup>. This suggests that there may be a therapeutic window for improving insulin sensitivity via PPAR $\gamma$ , using either a safe, low-dose treatment of a MEK inhibitor, a non-agonist ligand that blocks kinase accessibility to PPAR $\gamma$  S273, or both together. These data offer hope for resurrecting PPAR $\gamma$ -targeted therapeutics to improve whole-body insulin sensitivity.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 January; accepted 22 September 2014.

Published online 17 November 2014.

- Choi, J. H. *et al.* Anti-diabetic drugs inhibit obesity-linked phosphorylation of PPAR $\gamma$  by Cdk5. *Nature* **466**, 451–456 (2010).

- Choi, J. H. *et al.* Antidiabetic actions of a non-agonist PPAR $\gamma$  ligand blocking Cdk5-mediated phosphorylation. *Nature* **477**, 477–481 (2011).
- Yu, J. G. *et al.* The effect of thiazolidinediones on plasma adiponectin levels in normal, obese, and type 2 diabetic subjects. *Diabetes* **51**, 2968–2974 (2002).
- Tontonoz, P., Hu, E. & Spiegelman, B. M. Stimulation of adipogenesis in fibroblasts by PPAR $\gamma$ 2, a lipid-activated transcription factor. *Cell* **79**, 1147–1156 (1994).
- Chawla, A., Schwarz, E. J., Dimaculangan, D. D. & Lazar, M. A. Peroxisome proliferator-activated receptor (PPAR) gamma: adipose-predominant expression and induction early in adipocyte differentiation. *Endocrinology* **135**, 798–800 (1994).
- Saltiel, A. R. & Kahn, C. R. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414**, 799–806 (2001).
- Krüger, M. *et al.* Dissection of the insulin signaling pathway via quantitative phosphoproteomics. *Proc. Natl Acad. Sci. USA* **105**, 2451–2456 (2008).
- De Fea, K. & Roth, R. A. Modulation of insulin receptor substrate-1 tyrosine phosphorylation and function by mitogen-activated protein kinase. *J. Biol. Chem.* **272**, 31400–31406 (1997).
- Zheng, Y. *et al.* Improved insulin sensitivity by calorie restriction is associated with reduction of ERK and p70S6K activities in the liver of obese Zucker rats. *J. Endocrinol.* **203**, 337–347 (2009).
- Lazar, D. F. *et al.* Mitogen-activated protein kinase inhibition does not block the stimulation of glucose utilization by insulin. *J. Biol. Chem.* **270**, 20801–20807 (1995).
- Jiang, Z. Y. *et al.* Characterization of selective resistance to insulin signaling in the vasculature of obese Zucker (*fa/fa*) rats. *J. Clin. Invest.* **104**, 447–457 (1999).
- Jager, J. *et al.* Deficiency in the extracellular signal-regulated kinase 1 (ERK1) protects leptin-deficient mice from insulin resistance without affecting obesity. *Diabetologia* **54**, 180–189 (2011).
- Hawasli, A. H. *et al.* Cyclin-dependent kinase 5 governs learning and synaptic plasticity via control of NMDAR degradation. *Nature Neurosci.* **10**, 880–886 (2007).
- Eguchi, J. *et al.* Transcriptional control of adipose lipid handling by IRF4. *Cell Metab.* **13**, 249–259 (2011).
- Ohshima, T. *et al.* Targeted disruption of the cyclin-dependent kinase 5 gene results in abnormal corticogenesis, neuronal pathology and perinatal death. *Proc. Natl Acad. Sci. USA* **93**, 11173–11178 (1996).
- Hawasli, A. H. *et al.* Regulation of hippocampal and behavioral excitability by cyclin-dependent kinase 5. *PLoS ONE* **4**, e5808 (2009).
- Meijer, L. *et al.* Biochemical and cellular effects of roscovitine, a potent and selective inhibitor of the cyclin-dependent kinases cdc2, cdk2 and cdk5. *Eur. J. Biochem.* **243**, 527–536 (1997).

18. Bach, S. *et al.* Roscovitine targets, protein kinases and pyridoxal kinase. *J. Biol. Chem.* **280**, 31208–31219 (2005).
19. Shah, K., Liu, Y., Deirmengian, C. & Shokat, K. M. Engineering unnatural nucleotide specificity for Rous sarcoma virus tyrosine kinase to uniquely label its direct substrates. *Proc. Natl Acad. Sci. USA* **94**, 3565–3570 (1997).
20. Sun, K.-H., De Pablo, Y., Vincent, F. & Shah, K. Deregulated Cdk5 promotes oxidative stress and mitochondrial dysfunction. *J. Neurochem.* **107**, 265–278 (2008).
21. Tarricone, C. *et al.* Structure and regulation of the CDK5-p25(nck5a) complex. *Mol. Cell* **8**, 657–669 (2001).
22. Hu, E., Kim, J. B., Sarraf, P. & Spiegelman, B. M. Inhibition of adipogenesis through MAP kinase-mediated phosphorylation of PPAR $\gamma$ . *Science* **274**, 2100–2103 (1996).
23. Shao, D. *et al.* Interdomain communication regulating ligand binding by PPAR- $\gamma$ . *Nature* **396**, 377–380 (1998).
24. Rangwala, S. M. *et al.* Genetic modulation of PPAR $\gamma$  phosphorylation regulates insulin sensitivity. *Dev. Cell* **5**, 657–663 (2003).
25. Sharma, P. *et al.* Phosphorylation of MEK1 by cdk5/p35 down-regulates the mitogen-activated protein kinase pathway. *J. Biol. Chem.* **277**, 528–534 (2002).
26. Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40**, D261–D270 (2012).
27. Puigserver, P. *et al.* A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis. *Cell* **92**, 829–839 (1998).
28. Ohno, H., Shinoda, K., Spiegelman, B. M. & Kajimura, S. PPAR $\gamma$  agonists induce a white-to-brown fat conversion through stabilization of PRDM16 protein. *Cell Metab.* **15**, 395–404 (2012).
29. Flaherty, K. T. *et al.* Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations. *N. Engl. J. Med.* **367**, 1694–1703 (2012).
30. Solit, D. B. *et al.* BRAF mutation predicts sensitivity to MEK inhibition. *Nature* **439**, 358–362 (2006).

**Acknowledgements** We thank E. Rosen for providing us with the adiponectin Cre mice before their initial publication; members of the Spiegelman laboratory (Dana-Farber Cancer Institute) and D. Cohen (Brigham and Women's Hospital) for discussions; and C. Palmer and K. LeClair for reading the manuscript. B.M.S. acknowledges National Institutes of Health (NIH) grant DK31405. A.B. acknowledges NIH grant DK93638, the Harvard University Milton Fund, and the Harvard Digestive Disease Center, Core D.

**Author Contributions** A.B., B.M.S., F.E., S.G., J.P.C., M.J. and G.S. designed the experiments. A.B., D.B., F.E., J.C.P. and P.Z. performed the experiments. A.B., B.M.S. and F.E. wrote the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.B. ([abanks@research.bwh.harvard.edu](mailto:abanks@research.bwh.harvard.edu)) or B.M.S. ([bruce\\_spiegelman@dfci.harvard.edu](mailto:bruce_spiegelman@dfci.harvard.edu)).

## METHODS

**Animal experiments.** Animal experiments were performed with approval from the Institutional Animal Care and Use Committees of both Beth Israel Deaconess Medical Center and The Harvard Center for Comparative Medicine. Glucose and insulin tolerance tests were performed as described previously by researchers blinded to the genotype and treatment group to which each mouse belonged<sup>31</sup>. Glucose doses used were as follows: for mice maintained on a standard-diet, 2 g kg<sup>-1</sup>; high-fat diet, 1.5 g kg<sup>-1</sup>; *ob/ob* mice, 1 g kg<sup>-1</sup>. Male C57Bl/6 *J ob/ob* mice were purchased from Jackson Labs at 5–6 weeks of age and allowed to acclimate for 1–2 weeks before treatment. Cdk5<sup>Flox/Flox</sup> mice were provided by P. Greengard (Rockefeller University). Cdk5<sup>Flox/Flox</sup> control, adiponectin-Cre and the Cdk5-KO (Cdk5<sup>Flox/Flox</sup>; adiponectin-Cre) mice were healthy and viable, in contrast to mice with whole-body deletion of CDK5. Adiponectin-Cre mice were provided by E. Rosen (Beth Israel Deaconess Medical Center). Both strains were previously backcrossed to a C57Bl/6 background. For diet-induced obesity, male animals were fed on a high-fat (60%) diet (Research Diets, catalogue no. D12492i). For *in vivo* therapeutic assays, GSK1120212 (3 mg kg<sup>-1</sup>) or PD0325901 (refs 32–34) (10 mg kg<sup>-1</sup>) (Selleckchem) was administered by daily oral gavage for 5 days unless otherwise specified. Compounds were dissolved in dimethylsulphoxide and diluted into an aqueous 250- $\mu$ l dose containing 0.5% hypromellose and 2% Tween-80. High-fat-diet mice were switched to and maintained on a standard chow diet 48 h before the first dose of MEK inhibitors. Sample sizes were based not on power calculations but on the maximum number of mice that could be bred to within 2–3 weeks in age to maintain well-matched controls. For randomization of groups, mice were ranked according to body weight and alternated between vehicle and drug treatment groups. Mice were excluded from randomization if body weight was more than two standard deviations from the mean. Hyperinsulinaemic–euglycaemic clamp studies were performed as described previously<sup>2</sup>.

**Cell culture.** Cell culture of HEK 293 and F442A pre-adipocytes was performed as described: protein and RNA preparation, western blotting, quantitative real-time PCR with expression normalized to levels of TATA-binding protein (TBP)<sup>2</sup>. Cell lines were found free of mycoplasma before initiation of studies. Inhibition of analogue-sensitive CDK5 kinase (F80G) was performed by treating cells with 1NMPP1 (1-(1,1-dimethylethyl)-3-(1-naphthalenylmethyl)-1H-pyrazolo[3,4-d]pyrimidin-4-amine) (Cayman Chemical) at the indicated doses for 2 h.

**Plasmids.** The constitutively active ERK kinase, ERK-CA, is the product of a fusion of MEK1 with ERK2 (ERK2-MEK1-LA) and was provided by M. Cobb<sup>35</sup>. The inserts encoding human WT-MEK2 and CA-MEK2 (S222D and S226D) from Addgene (catalogue nos 40776 and 29580) were cloned into Flag-pCDNA3.1. The F80G mutation of haemagglutinin-tagged Cdk5 was introduced using the Quikchange XL Site Directed Mutagenesis Kit (Agilent). This insert was cloned into the pMSCV backbone retrovirus for stable infection of F442A pre-adipocytes.

**Antibodies.** Antibodies were obtained from Cell Signaling Technology (anti-phospho-ERK1/2 (catalogue no. 9101), anti-ERK (catalogue no. 4695), anti-phospho p38 (catalogue no. 4511), anti-p38 (catalogue no. 9212), anti-phospho JNK (catalogue no. 4671), anti-JNK (catalogue no. 9252), anti-phospho AKT (catalogue no. 13038) and total AKT (catalogue no. 9272)), Millipore (anti-phospho-S112 PPAR $\gamma$  (catalogue no. 04-816)) and Santa Cruz Biotechnology (anti-PPAR $\gamma$  (catalogue no. sc-7273) and anti-phospho-394-MEK2 (catalogue no. sc-101734)). Antibodies against anti-phospho-S273 PPAR $\gamma$  were generated as described previously<sup>1</sup>.

**Indirect calorimetry.** Energy expenditure, O<sub>2</sub> consumption, CO<sub>2</sub> production, respiratory exchange ratio, total locomotor activity and food intake measurements were made with a 16-cage Columbus Instruments Oxymax Comprehensive Lab Animal Monitoring System at ambient room temperature (21–23 °C). Whole-body composition was assessed with an EchoMRI 3-in-1 on conscious mice both before and after calorimetry. Because body weight and body composition were unchanged between WT and Cdk5-KO mice, data were analysed by analysis of variance (ANOVA).

**Mass spectrometry.** Enrichment using ActivX ATP probes (Thermo) combined with phosphopeptide enrichment were used to profile kinases<sup>36</sup>. In brief, tissue extracts were incubated in the presence of non-hydrolysable ATP analogues coupled to a desthiobiotinylated tag. These small-molecule probes are designed to covalently attach to ATPases, including protein kinases. Peptides were then labelled with TMT isobaric tags and the resulting mass spectra were analysed quantitatively. The method was recently adapted to include an additional phosphopeptide enrichment step, thereby improving the identification and quantification of protein kinase activities<sup>37</sup>. The methods are similar to those described in ref. 37. Mouse tissue homogenate from three WT and three Cdk5-KO mice was subjected to gel filtration in spin columns (Zeba; Pierce) in accordance with the manufacturer's instructions to remove endogenous ATP, ADP and small molecules and then diluted with reaction buffer (25 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% Nonidet P40, 5% glycerol) to a final protein concentration of 2 mg ml<sup>-1</sup>. Protease inhibitors (1  $\times$  'complete'; Roche 04693132001) and phosphatase inhibitors (final

concentration 2 mM imidazole, 1 mM sodium fluoride, 1.15 mM sodium molybdate, 4 mM sodium tartrate dehydrate, 1 mM  $\beta$ -glycerophosphate, 50  $\mu$ M phenylarsine) were added along with MnCl<sub>2</sub> to a final concentration of 10  $\mu$ M. Lysates were incubated with ActivX ATP probes for 10 min at room temperature at a concentration of 20  $\mu$ M. The reaction was quenched with 8 M urea, reduced with dithiothreitol (5 mM final concentration) and then alkylated with iodoacetamide (15 mM final concentration). The solution was then subjected again to gel filtration (Zeba; Pierce). Streptavidin was then added to the lysate to capture the undigested, desthiobiotinylated proteins and kinases. After extensive washing (lysis buffer containing 6 M urea (five times with 100  $\mu$ l), then 50 mM HEPES (five times with 100  $\mu$ l)), the captured proteins were subjected to on-bead digestion with trypsin (5 ng  $\mu$ l<sup>-1</sup>) for 4 h at 37 °C in a tandem-mass-tag-compatible buffer (50 mM HEPES pH 7.4, 0.5 M guanidinium chloride). After digestion, the resulting peptides were extracted and the beads were washed with 50 mM HEPES (twice with 50  $\mu$ l) and these washes were added to the peptide mixture. Acetonitrile was added to the peptide mixture to a final concentration of 30%, and the peptides were subjected to Tandem mass tags (TMT; Thermo Scientific) labelling. For labelling, 0.8 mg of each TMT reagent (126, 127, 128, 129, 130 and 131) was resuspended in 40  $\mu$ l of anhydrous acetonitrile. Peptides were resuspended in 17.5  $\mu$ l of 50 mM HEPES pH 8.5 and 5  $\mu$ l of acetonitrile, to which 2.5  $\mu$ l TMT reagent was added. The TMT labelling reaction was performed at room temperature for 1 h, and individual labelling reactions were quenched by the addition of 3  $\mu$ l of 5% hydroxylamine. The six samples were then combined and desalted with StageTips<sup>38</sup>.

For phosphopeptide enrichment, peptides were resuspended in 100  $\mu$ l of binding buffer (2 M lactic acid, 50% acetonitrile). Phosphopeptides were enriched with TiO<sub>2</sub> as described<sup>39</sup>. TiO<sub>2</sub> resin (600  $\mu$ g; GL Sciences) was prepared by washing twice with 200  $\mu$ l of binding buffer and was then added to the peptides in binding buffer and incubated for 1 h at room temperature. After incubation, beads were recovered by centrifugation (4 min at 200g) and washed with binding buffer (five times with 200  $\mu$ l). Bound phosphopeptides were then eluted with 50 mM K<sub>2</sub>HPO<sub>4</sub> pH 10 (three times with 20  $\mu$ l) and further purified with StageTip. The purified phosphopeptides were resuspended in 8  $\mu$ l of 5% formic acid, and 4  $\mu$ l was injected and analysed by LC-MS3.

In the Cdk5 experiment, tissue homogenate was pretreated with Cdk5/p35 (Millipore) for 10 min with the indicated amounts before being processed with the above protocol. ActivX ATP probes, high-capacity binding streptavidin and TMT were obtained from Thermo Scientific; modified trypsin was obtained from Promega; and TiO<sub>2</sub> beads were obtained from GL Sciences. SepPak C<sub>18</sub> solid-phase extraction cartridges were purchased from Waters Corporation.

**Liquid chromatography (LC) and mass spectrometry (MS) analysis.** LC-MS/MS analysis was performed on an LTQ Orbitrap Velos or an LTQ Orbitrap Elite mass spectrometer (Thermo-Fisher Scientific) linked to an Accela 600 quaternary LC pump (Thermo) and a Famos autosampler (LC Packings). Flow rates of 300 nl min<sup>-1</sup> over the column were achieved by using a flow-split method. A hand-pulled fused silica microcapillary column (125  $\mu$ m  $\times$  18 cm) was used for peptide separation. The column was first packed with about 0.5 cm of Magic C4 resin (particle size 5  $\mu$ m, pore size 100 Å; Michrom Bioresources) and then with 18 cm of Maccel C<sub>18</sub> AQ resin (particle size 3  $\mu$ m, pore size 200 Å; Nest Group). The total LC-MS run length for each sample was 180 min and consisted of a 150-min gradient from 3% to 33% acetonitrile in 0.125% formic acid. A recently developed MS3 method was used to overcome the interference problem in the acquisition of TMT data<sup>40</sup>. In brief, a high-resolution MS1 scan in the Orbitrap (300–1500 *m/z*, 60k resolution; automatic gain control (AGC) 10<sup>6</sup>) was collected from which the top ten precursors were selected for MS2 analysis followed by MS3 analysis. The MS2 scan was performed in the quadrupole ion trap (collision-induced dissociation, AGC 2  $\times$  10<sup>3</sup>, normalized collision energy 35, maximum injection time 100 ms) and the MS3 scan was analysed in the Orbitrap (HCD, 30k resolution, maximum AGC 1.5  $\times$  10<sup>5</sup>, maximum injection time 250 ms, normalized collision energy 50). Multiple fragment ions from each MS2 spectrum were selected for MS3 analysis using isolation waveforms with multiple frequency notches<sup>41</sup>.

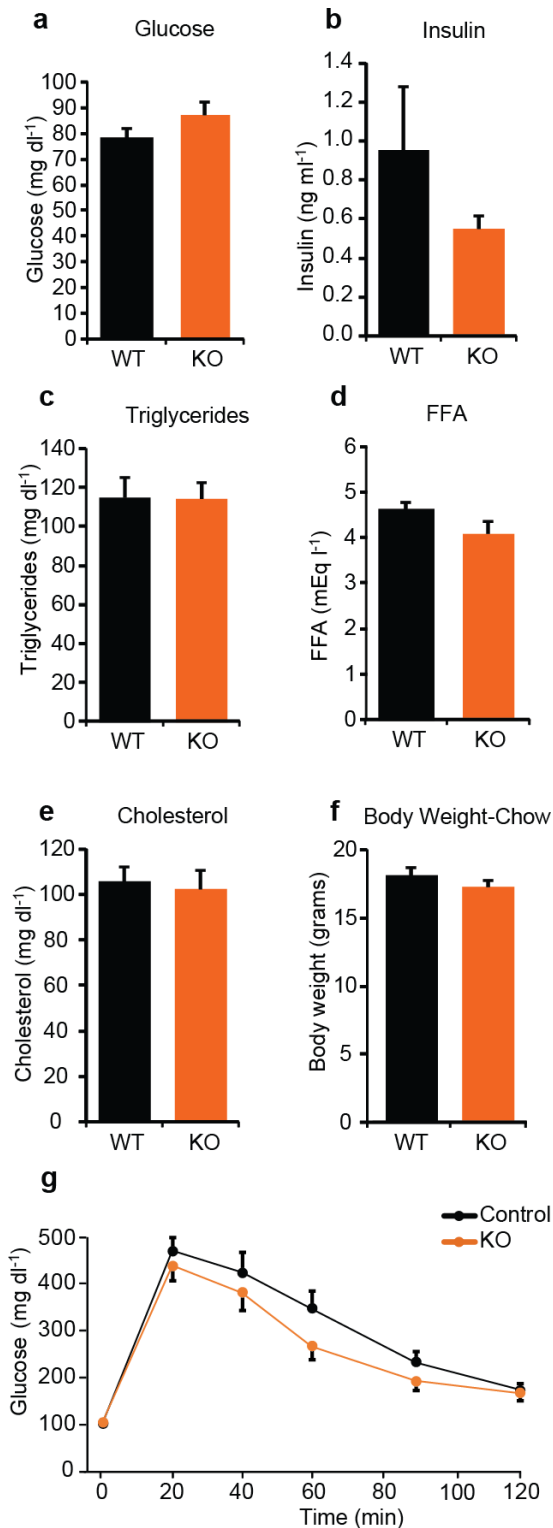
**Data analysis.** Statistical data analysis was performed with GraphPad Prism and Microsoft Excel, including Daniel's XL Toolbox Add-In. Unless otherwise specified, analyses were by one-tailed Student's *t*-test. Sequence alignment was performed with a modified Clustal W algorithm using Vector NTI AlignX. Mass spectrometry data were processed using an in-house software pipeline<sup>42</sup>. Raw files were converted to mzXML files and searched using the Sequest algorithm<sup>43</sup> against a composite database containing sequences from the mouse uniprot database in forward and reverse orientations as well as the sequences of common contaminating proteins (for example trypsin). Database searching matched MS/MS spectra with fully tryptic peptides from this composite database with a 20 p.p.m. precursor ion and a product ion tolerance of 1 Da. Carbamidomethylation of cysteine residues (+57.02146 Da) and TMT tags on peptide amino termini and lysines (+229.162932 Da) were set as



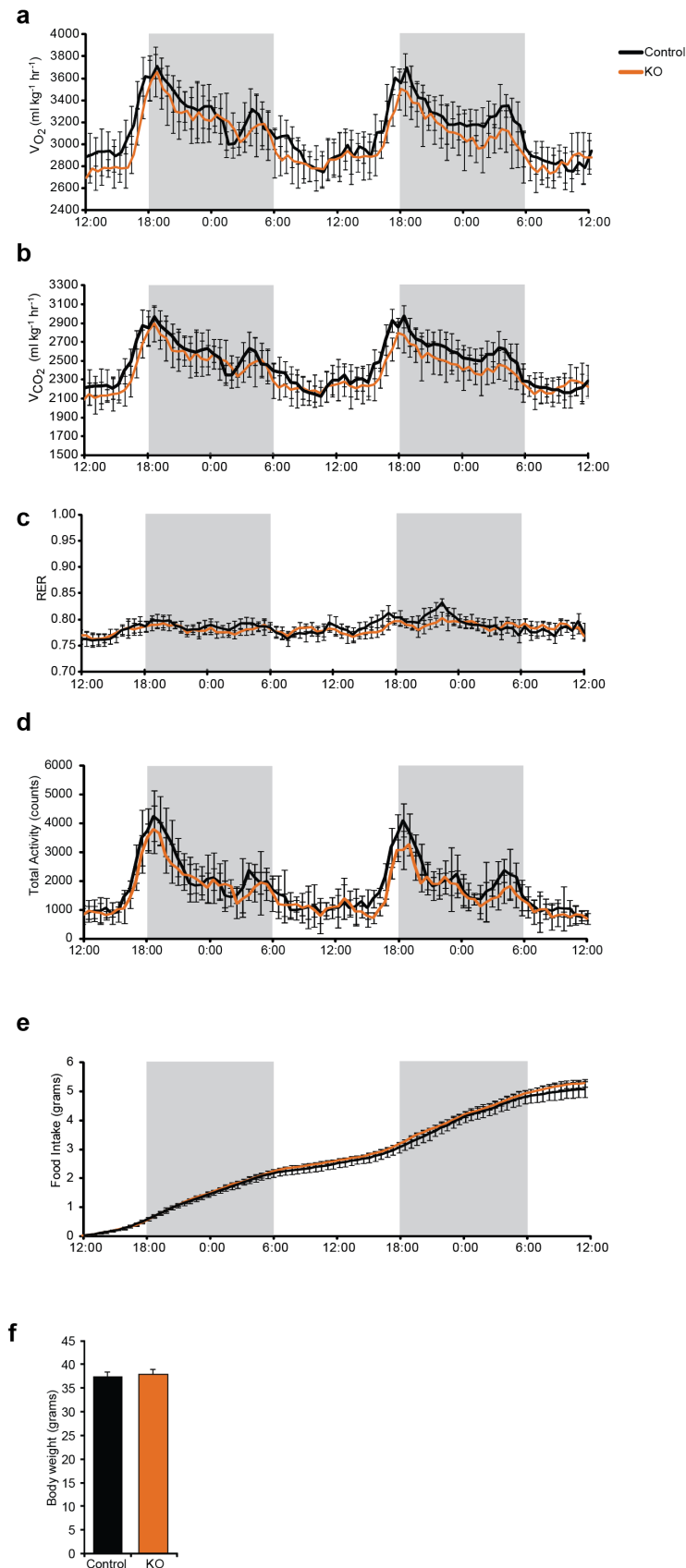
static modifications. Variable modifications of oxidation of methionine residues (+15.99492 Da) and phosphorylation (+79.966330 Da) on serine, threonine and tyrosine residues were used. The data were filtered to a false discovery rate of less than 1% based on the target-decoy database approach at both the peptide and protein levels<sup>44</sup>. Linear discriminant analysis was performed to generate a classifier to distinguish between correct and incorrect MS2 spectra assignments based on the following parameters: XCorr,  $\Delta$ Cn, peptide ion mass accuracy, charge state and peptide length, as described<sup>42</sup>. Peptides were then assembled into proteins that were scored probabilistically and further filtered to a protein-level false discovery rate of about 1%. Peptide quantification using TMT reporter ion intensity was performed using in-house software, as described<sup>42</sup>. In brief, a 0.06  $m/z$  window around the theoretical  $m/z$  value of each reporter ion was scanned for ions, and the intensity of the signal nearest to the theoretical  $m/z$  value was recorded. The intensities of the reporter ions were adjusted to account for isotopic impurities in each TMT variant (as provided by the manufacturer). For comparisons, the whole data sets were filtered to a 1% false discovery rate, and proteins and phosphosites were then quantified by summing reporter ion counts for all the peptide-spectral matches. Filtering was performed to remove poor-quality MS3 spectra in a manner similar to that described previously<sup>41</sup>. Protein quantification values were exported for further analysis in Excel or Matlab. Hierarchical clustering was performed using Matlab.

**In vitro, MS-based kinase activity assay.** PPAR $\gamma$  (1  $\mu$ g; Active Motif) was incubated with 50 ng of recombinant kinases MEK2, ERK2 (SignalChem) and Cdk5/p35 (Millipore) and kinase reaction buffer containing 25 mM Tris-HCl pH 7.5, 5 mM ATP, 7.5 mM MgCl<sub>2</sub>, 0.2 mM EGTA, 7.5 mM  $\beta$ -glycerophosphate, 0.1 mM Na<sub>3</sub>VO<sub>4</sub> and 0.1 mM dithiothreitol in a final reaction volume of 50  $\mu$ l. All proteins corresponded to the human sequences. After incubation for 45 min at room temperature, the reaction mixture was subjected to both LysC and trypsin digestion (individual reactions). After purification, the samples were then analysed by LC-MS/MS in a similar manner to that described above but without the additional MS3 dimension.

31. Banks, A. S. *et al.* Dissociation of the glucose and lipid regulatory functions of FoxO1 by targeted knockin of acetylation-defective alleles in mice. *Cell Metab.* **14**, 587–597 (2011).
32. Barrett, S. D. *et al.* The discovery of the benzhydroxamate MEK inhibitors CI-1040 and PD 0325901. *Bioorg. Med. Chem. Lett.* **18**, 6501–6504 (2008).
33. Albeck, J. G., Mills, G. B. & Brugge, J. S. Frequency-modulated pulses of ERK activity transmit quantitative proliferation signals. *Mol. Cell* **49**, 249–261 (2013).
34. Lau, K. S. *et al.* In vivo systems analysis identifies spatial and temporal aspects of the modulation of TNF- $\alpha$ -induced apoptosis and proliferation by MAPKs. *Sci. Signal.* **4**, ra16 (2011).
35. Robinson, M. J., Stippes, S. A., Goldsmith, E., White, M. A. & Cobb, M. H. A constitutively active and nuclear form of the MAP kinase ERK2 is sufficient for neurite outgrowth and cell transformation. *Curr. Biol.* **8**, 1141–1150 (1998).
36. Patricelli, M. P. *et al.* In situ kinase profiling reveals functionally relevant properties of native kinases. *Chem. Biol.* **18**, 699–710 (2011).
37. McAllister, F. E. *et al.* Mass spectrometry based method to increase throughput for kinome analyses using ATP probes. *Anal. Chem.* **85**, 4666–4674 (2013).
38. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature Protocols* **2**, 1896–1906 (2007).
39. Kettenbach, A. N. & Gerber, S. A. Rapid and reproducible single-stage phosphopeptide enrichment of complex peptide mixtures: application to general and phosphotyrosine-specific phosphoproteomics experiments. *Anal. Chem.* **83**, 7635–7644 (2011).
40. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods* **8**, 937–940 (2011).
41. McAllister, G. C. *et al.* Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal. Chem.* **84**, 7469–7478 (2012).
42. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010).
43. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
44. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207–214 (2007).

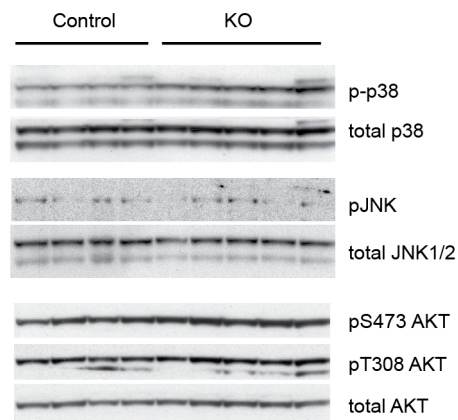


**Extended Data Figure 1 | Metabolic profiling of adipose-specific Cdk5-KO mice on a standard chow diet.** **a–e**, Fasting plasma levels of glucose (**a**), insulin (**b**), total triacylglycerols (**c**), free fatty acids (FFA) (**d**) and total cholesterol (**e**) ( $n = 16$  (control) and 17 (KO)). **f, g**, Body weights (**f**) and intraperitoneal glucose tolerance test (**g**). Mice were 12 weeks of age ( $n = 14$  (control) and 11 (KO)). No significant differences were observed. Error bars indicate s.e.m.



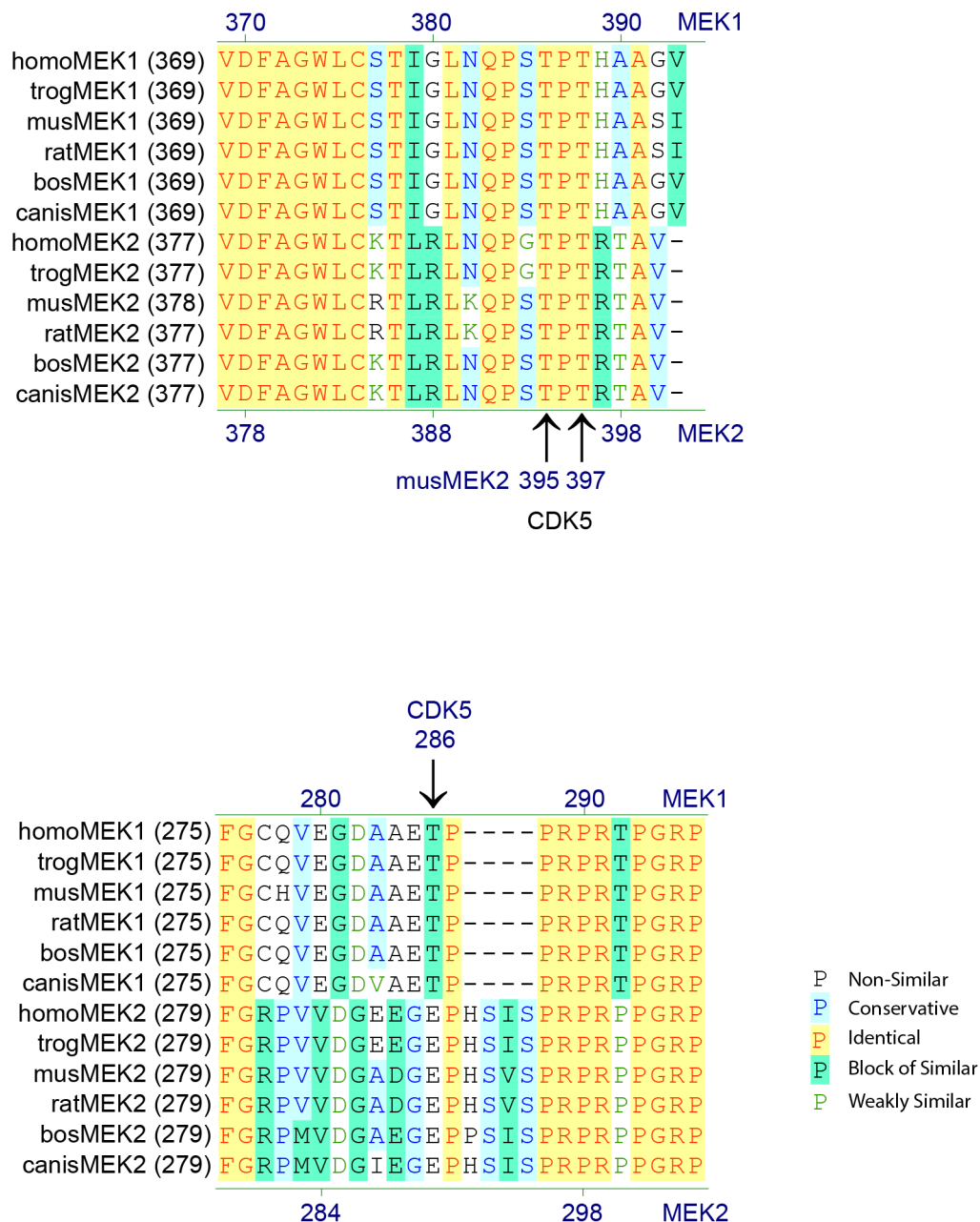
**Extended Data Figure 2 | Energy homeostasis of adipose-specific Cdk5-KO mice maintained on a high-fat diet.** **a–f**, After a 48-h acclimatization period, singly housed mice were monitored for oxygen consumption ( $V_{O_2}$ ) (**a**), carbon dioxide production ( $V_{CO_2}$ ) (**b**), respiratory exchange ratio (RER) (**c**),

ambulatory locomotor activity (**d**), cumulative food intake (**e**) and body weights (**f**) ( $n = 8$  per group). Shaded areas signify the dark phase of the light cycle. No significant differences were observed. Error bars indicate s.e.m.



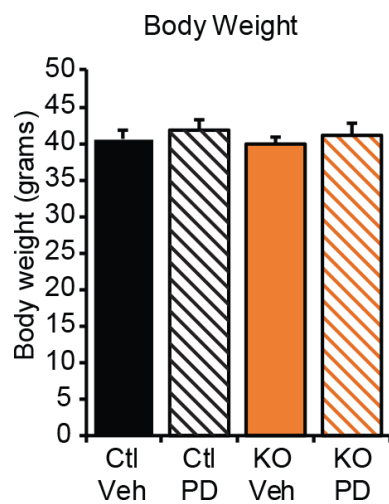
**Extended Data Figure 3 | Activity of alternative kinases in adipose tissue from Cdk5-KO mice.** Brown adipose tissue protein lysates from mice maintained on a high-fat diet for 12 weeks. Blotting for phospho-p38, phospho-JNK and phospho-S473 and pT308 AKT was performed before loading for total protein amounts.



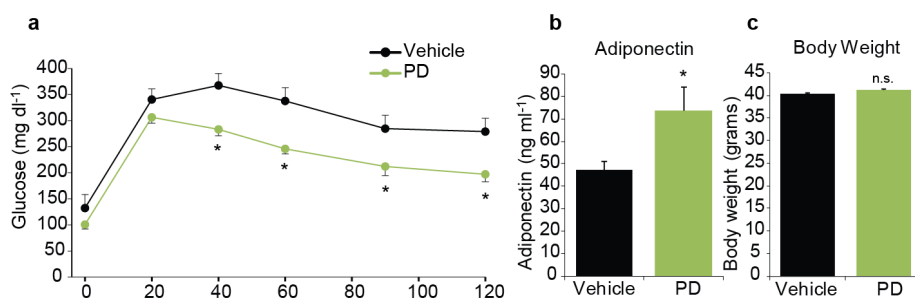


**Extended Data Figure 4 | Conservation of the sites on MEK2 phosphorylated by Cdk5.** Mouse MEK2 T395/T397 corresponds to human MEK2 T394/T396. These sites share identity with MEK1 T386/T388 in both humans and mouse. Cdk5 has previously been shown to phosphorylate MEK1

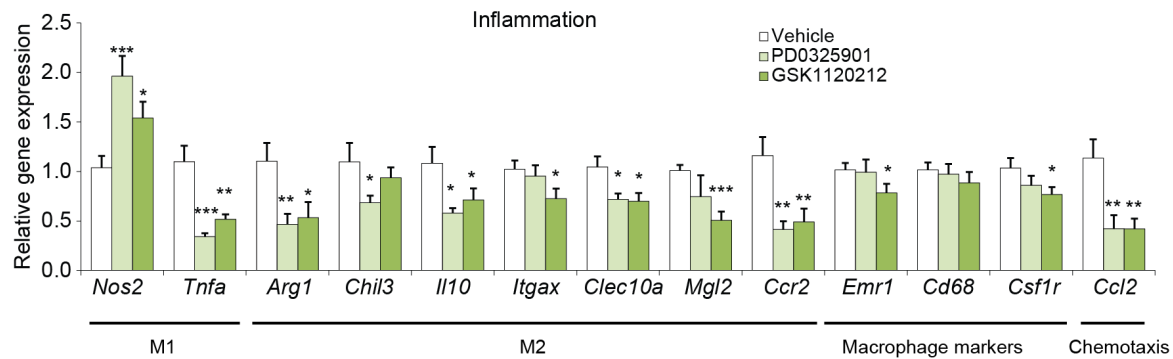
at T286, a site not shared with MEK2. ERK has been shown to phosphorylate MEK1 T386 and contribute to regulation of kinase activity<sup>31</sup>. Homo, *Homo sapiens*; trog, *Pan troglodytes*; mus, *Mus musculus*; rat, *Rattus norvegicus*; bos, *Bos taurus*; canis, *Canis lupus familiaris*.



**Extended Data Figure 5 | Body weight of control and of adipose-specific Cdk5-KO mice maintained on a high-fat diet after treatment with PD0325901.** Treatment similar to that in Fig. 4a–c. The body weights are not significantly different by ANOVA. Error bars indicate s.e.m.



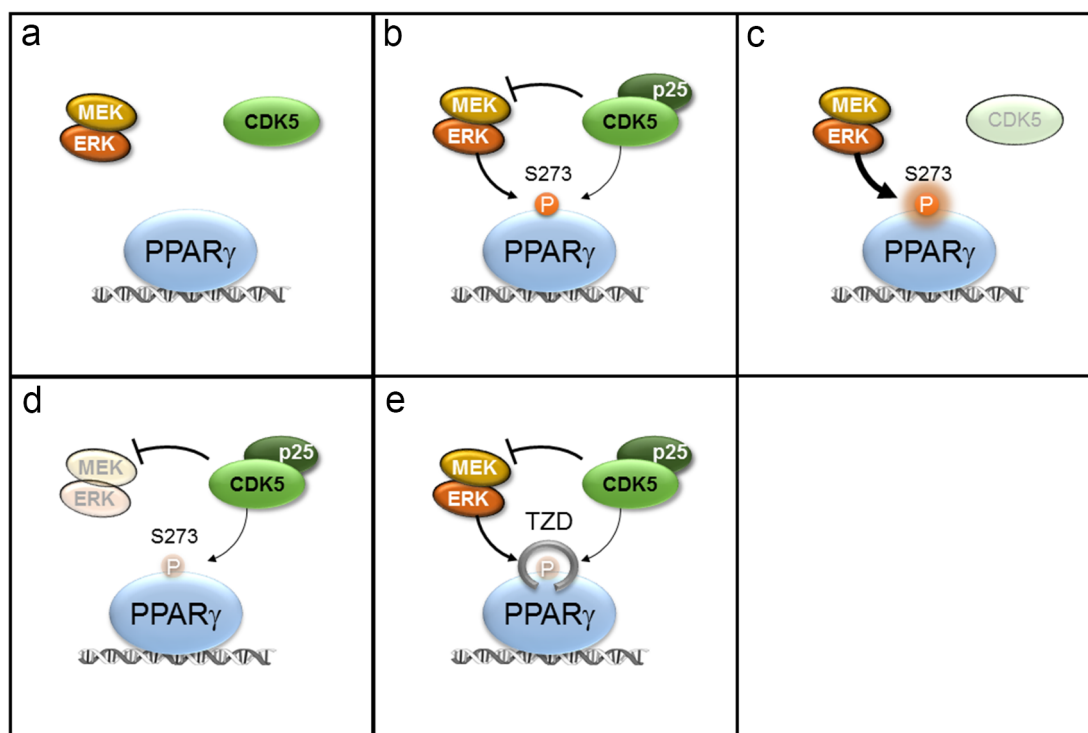
**Extended Data Figure 6** | Effects of PD0325901 treatment on *ob/ob* mice. **a–c**, Glucose tolerance test (**a**), adiponectin levels (**b**) and body weights (**c**) of *ob/ob* mice treated with PD0325901 ( $n = 7$  (vehicle) and 8 (PD)). \* $P \leq 0.05$  by Student's *t*-test. Error bars indicate s.e.m.



**Extended Data Figure 7 | Inflammatory markers in epididymal white adipose tissue from *ob/ob* mice treated with MEK inhibitors.** Gene expression analysis was performed on M1 macrophage markers *Nos2* and tumour necrosis factor- $\alpha$  (TNF- $\alpha$ ); M2 macrophage markers *Arg1*, *Chil3*, *Il10*,

*Itgax*, *Clec10a*/*Mgl1* and *Mgl2*; chemotactic ligand *Ccl2* and receptor *Ccr2*; and macrophage surface markers *Emr1*, *Cd68* and *Csf1r* ( $n = 7$  or  $8$  mice per group as in Fig. 5f, h). Gene expression was analysed by ANOVA. Error bars indicate s.e.m. \* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ .





**Extended Data Figure 8 | Schematic model of PPAR $\gamma$  regulation at S273.**

**a**, In the lean state, PPAR $\gamma$  is not phosphorylated. **b**, In the obese state, S273 phosphorylation is driven by both Cdk5 and ERK with CDK5 repressing MEK and ERK activity. **c**, Cdk5-KO results in derepression of MEK and ERK kinases

and increased phosphorylation of S273 PPAR $\gamma$ . **d**, MEK inhibition markedly decreases S273 PPAR $\gamma$  phosphorylation. **e**, PPAR $\gamma$  ligands, including the thiazolidinediones, block the accessibility of S273 PPAR $\gamma$  by either ERK or CDK5 kinases.

# Subnanometre-resolution electron cryomicroscopy structure of a heterodimeric ABC exporter

JungMin Kim<sup>1\*</sup>, Shenping Wu<sup>2\*</sup>, Thomas M. Tomasiak<sup>2\*</sup>, Claudia Mergel<sup>3</sup>, Michael B. Winter<sup>1</sup>, Sebastian B. Stiller<sup>3</sup>, Yaneth Robles-Colmanares<sup>2</sup>, Robert M. Stroud<sup>1,2</sup>, Robert Tampé<sup>3,4</sup>, Charles S. Craik<sup>1</sup> & Yifan Cheng<sup>2</sup>

ATP-binding cassette (ABC) transporters translocate substrates across cell membranes, using energy harnessed from ATP binding and hydrolysis at their nucleotide-binding domains<sup>1,2</sup>. ABC exporters are present both in prokaryotes and eukaryotes, with examples implicated in multidrug resistance of pathogens and cancer cells, as well as in many human diseases<sup>3,4</sup>. TmrAB is a heterodimeric ABC exporter from the thermophilic Gram-negative eubacterium *Thermus thermophilus*; it is homologous to various multidrug transporters and contains one degenerate site with a non-catalytic residue next to the Walker B motif<sup>5</sup>. Here we report a subnanometre-resolution structure of detergent-solubilized TmrAB in a nucleotide-free, inward-facing conformation by single-particle electron cryomicroscopy. The reconstructions clearly resolve characteristic features of ABC transporters, including helices in the transmembrane domain and nucleotide-binding domains. A cavity in the transmembrane domain is accessible laterally from the cytoplasmic side of the membrane as well as from the cytoplasm, indicating that the transporter lies in an inward-facing open conformation. The two nucleotide-binding domains remain in contact via their carboxy-terminal helices. Furthermore, comparison between our structure and the crystal structures of other ABC transporters suggests a possible trajectory of conformational changes that involves a sliding and rotating motion between the two nucleotide-binding domains during the transition from the inward-facing to outward-facing conformations.

ABC transporters use ATP binding and hydrolysis to drive substrate translocation across a membrane. Many members of the ABC exporter family have varying selectivity and transport substrates from inside to outside the cell, a property thought to allow them to facilitate export of xenobiotics such as drugs and toxins<sup>6,7</sup>. TmrAB has similar features to multidrug transporters, including transport of Hoechst 33342 dye and competitive inhibition by verapamil, which suggests a common mechanism for transport<sup>5</sup>. It is composed of two homologous subunits, TmrA and TmrB, arranged with pseudo-two-fold symmetry with a combined molecular mass of ~135 kilodaltons (kDa). Each subunit has a six-helix transmembrane domain (TMD) and a cytoplasmic nucleotide-binding domain (NBD). TmrAB has two ATP binding sites, each formed between both NBDs. However, only one site is an active ATPase consensus site capable of ATP hydrolysis. The degenerate ('inactive') site has a non-canonical aspartate residue next to the Walker B motif contributed by one NBD and non-canonical residues from the ABC signature motif of the other NBD<sup>5</sup>.

Despite several crystal structures of ABC exporters representing various states along the transport cycle, there are competing models of how these states are functionally connected in a physiological setting<sup>8–10</sup>. Single-particle electron cryomicroscopy (cryo-EM) holds the promise of offering structural information complementary to X-ray crystallography, especially for conformational states that may be difficult to access within the confines of a crystal lattice. However, structure determination by single-particle cryo-EM is favoured by a relatively large molecular mass

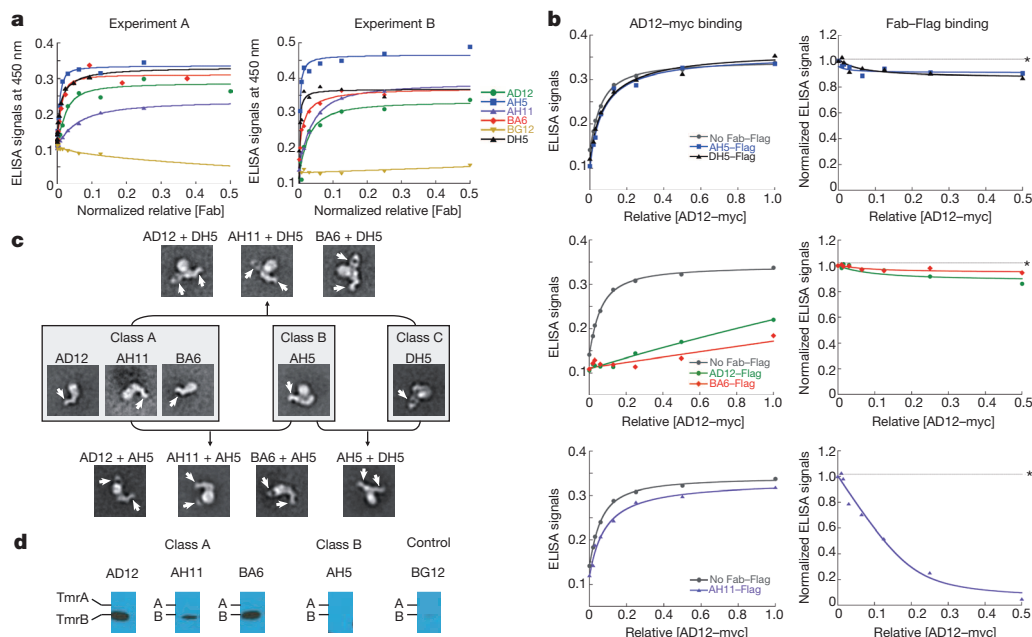
and higher symmetry. Although recent technological breakthroughs have enabled determination of the first atomic structures of homo-tetrameric ion channels of ~300 kDa (ref. 11), determining a high-resolution structure of TmrAB still represents a major challenge, owing to its smaller size and pseudo-symmetric organization<sup>12</sup>.

Here, we used fragment antigen-binding (Fab) domains to overcome these challenges. A Fab that forms a stable complex with TmrAB has a number of advantages for structure determination by single-particle cryo-EM<sup>13</sup>. Furthermore, conformational-specific synthetic Fabs stabilize particles in a specific functional state<sup>14</sup>. Following our established procedure for Fab selection<sup>15</sup>, five Fabs were identified from a human naive B-cell Fab phage-displayed library using *n*-dodecyl- $\beta$ -D-maltopyranoside ( $\beta$ -DDM)-solubilized TmrAB as the antigen. Fab binding was validated using a qualitative enzyme-linked immunosorbent assay (ELISA) screen<sup>15</sup> (Fig. 1a and Extended Data Fig. 1a). The Fabs were further characterized by competitive ELISA analysis to establish whether they have overlapping or independent epitopes (see Methods and Fig. 1b). AD12, BA6, and AH11 ('class A') were found to have overlapping epitopes, inhibiting the binding of one another, whereas AH5 ('class B') and DH5 ('class C') were found to have unique epitopes. Rigidity of the complexes was assessed by negative-stain electron microscopy two-dimensional class averages. All TmrAB–Fab complexes yielded two-dimensional class averages that show characteristic features of Fabs (Fig. 1c and Extended Data Fig. 1d), suggesting that these Fabs form sufficiently rigid complexes with TmrAB. TmrAB complexes that clearly show two Fabs (Fig. 1c) confirm that those Fabs bind to distinct sites. Furthermore, ELISA and negative-stain electron microscopy demonstrated that AH5 and BA6 display the highest relative affinities (Extended Data Fig. 1b–d). Thus, these Fabs are preferred candidates for structure determination of TmrAB by cryo-EM. To gain more insight into the Fab-binding properties, AH5 and class A Fabs were used in an immunoblotting assay with denatured TmrAB. AH5, despite showing the highest affinity (Extended Data Fig. 1d), did not bind the denatured transporter, whereas all three class A Fabs, including BA6, bound to denatured TmrB (Fig. 1d). These data suggest that AH5 recognizes a three-dimensional epitope and that BA6 recognizes a linear epitope(s) within the TmrB sequence.

Detergent-solubilized TmrAB–AH5 complex was purified (Extended Data Fig. 2a). Frozen-hydrated TmrAB–AH5 particles were imaged using a direct electron-detection camera, K2 Summit, following newly implemented procedures<sup>16</sup>. Two-dimensional class averages of the TmrAB–AH5 complex show landmark features of both TmrAB and Fab (Extended Data Fig. 2b–e). We determined a three-dimensional reconstruction of TmrAB–AH5 to a resolution of 8.2 Å using gold-standard Fourier shell correlation = 0.143 criterion (Extended Data Fig. 2f, g)<sup>17</sup>. Local resolution estimation<sup>18</sup> suggests most regions of the density map have a ~6 Å resolution (Extended Data Fig. 2h, i). The density map is of sufficient quality to resolve the secondary structure features clearly, including all helices in the TMDs and NBDs (Fig. 2 and Extended Data Fig. 3). Such

<sup>1</sup>Department of Pharmaceutical Chemistry, University of California San Francisco, 600 16th Street, San Francisco, California 94158, USA. <sup>2</sup>Department of Biochemistry and Biophysics, University of California San Francisco, 600 16th Street, San Francisco, California 94158, USA. <sup>3</sup>Institute of Biochemistry, Biocenter, Goethe-University Frankfurt, Max-von-Laue-Strasse 9, D-60438 Frankfurt am Main, Germany. <sup>4</sup>Cluster of Excellence – Macromolecular Complexes, Goethe-University Frankfurt, Max-von-Laue-Strasse 9, D-60438 Frankfurt am Main, Germany.

\*These authors contributed equally to this work.

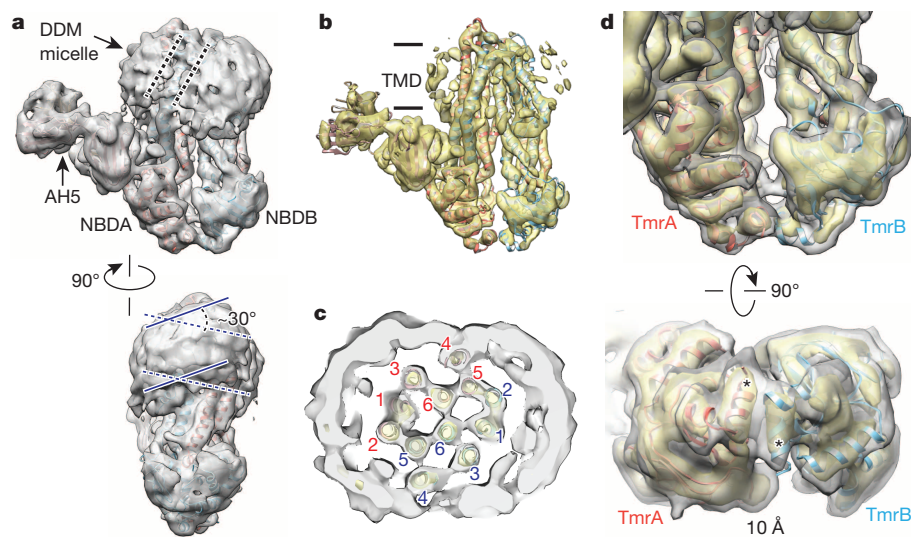


**Figure 1 | TmrAB Fab characterization.** **a**, Qualitative ELISA to assess Fab binding to TmrAB. All Fabs showed binding to TmrAB except BG12 in both experiments A and B, where independently prepared Fab samples were used. Expression levels were assessed by immunoblotting and normalized (Extended Data Fig. 1a). **b**, Representative competitive ELISA between AD12-myc and Fab-Flag against TmrAB. AD12-myc binding was not affected by the presence of AH5-Flag or DH5-Flag (top left). AH5-Flag and DH5-Flag maintained near-maximum binding (\*) at all AD12-myc concentrations (top right). AD12-myc binding was almost abolished in the presence of AD12-Flag or BA6-Flag (middle left). AD12-Flag and BA6-Flag maintained near-maximum binding (\*) at all AD12-myc concentrations (middle right).

AD12-myc binding was not significantly affected in the presence of AH11-Flag (bottom left). AH11-Flag binding decreased as AD12-myc concentrations increased shown (bottom right). Uninhibited binding of AD12-myc and Fab-Flag suggests independent binding between AD12 and AH5 or DH5. Inhibited binding of either AD12-myc or Fab-Flag suggests overlapping epitopes between AD12 and BA6 or AH11. **c**, Representative negative-stain two-dimensional class averages of TmrAB-Fab complexes. **d**, Immunoblotting of TmrAB, using Fab-Flag. Class A Fabs recognized the denatured form of TmrB, and AH5 did not recognize the denatured form of either strand. BG12 was used as a non-binder control, which did not detect either strand significantly.

well-defined structural features suggest that TmrAB as visualized here adapts a defined conformation of an ABC exporter with well-ordered domain architecture. Furthermore, density corresponding to the detergent

micelle defines a distorted ellipsoidal structure around the TMDs. The characteristic dumb-bell-shaped Fab density is also clearly defined, validating the correctness of the three-dimensional reconstruction.



**Figure 2 | Three-dimensional reconstruction of TmrAB-AH5 at subnanometre resolution.** **a**, Cryo-EM density map of the TmrAB-AH5 complex filtered to a resolution of 8.2 Å. The atomic model of TmrAB and an atomic structure of a Fab (Protein Data Bank accession number 1M71) are docked into the density map. The map shows two NBDs, a bi-lobed DDM micelle, which is separated by TM4 (marked by two dotted lines in the top view), and well-defined AH5 density. In the bottom view, two parallel solid lines and dotted lines indicate orientations of the front and back halves of the

micelle respectively. The two halves are tilted by  $\sim 30^\circ$  with respect to each other. **b**, The density map at a higher contour level shows clearly resolved transmembrane helices. **c**, A cross-section view through the TMDs shows well-resolved transmembrane helices labelled red and blue for TmrA and TmrB respectively. **d**, NBDs of TmrAB in two different views. C-terminal helices (\*) of the two NBDs are in close proximity depicted both in the density map and the docked atomic model.



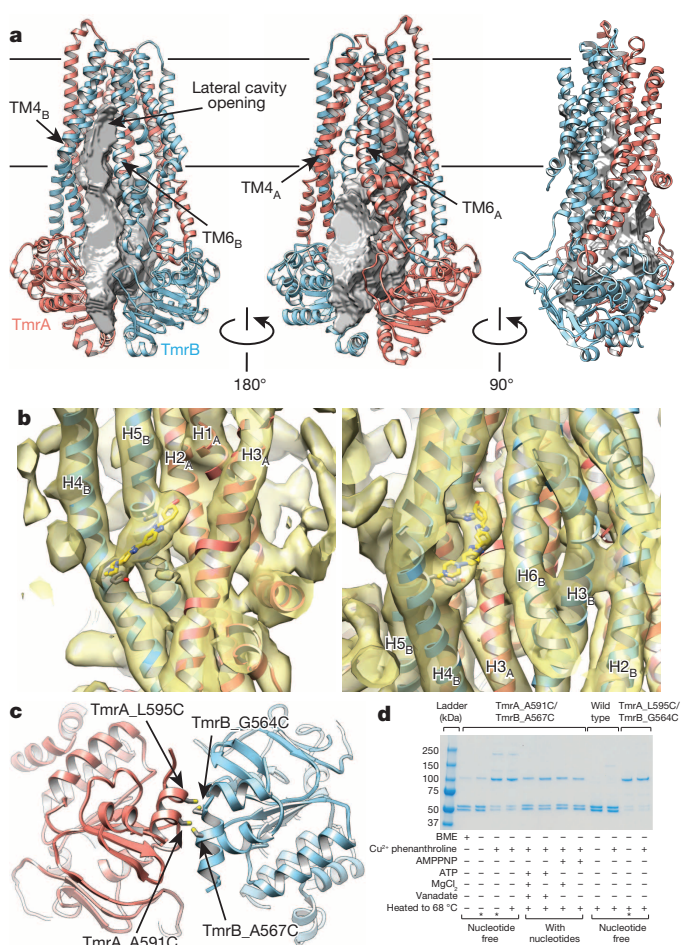
To confirm that the conformation of TmrAB is not influenced by AH5 binding, we determined two other three-dimensional reconstructions: TmrAB alone and in complex with BA6. The three-dimensional reconstruction of the TmrAB–BA6 complex was determined to a resolution of 9.4 Å (Extended Data Fig. 4) using a phosphor scintillator-based complementary metal-oxide-semiconductor (CMOS) camera. At this resolution, the majority of the TmrAB helices are resolved, and the BA6 Fab density has the expected shape. The three-dimensional reconstruction of TmrAB alone was determined using the same CMOS camera to 10 Å resolution (Extended Data Fig. 5). The densities of TmrAB in all reconstructions overlap with each other in a unique orientation as measured by local cross correlation (Extended Data Fig. 6), confirming that TmrAB is in a native conformation that is not induced by Fab binding. These experiments also demonstrated that both Fab incorporation and utilization of the direct electron-detection camera improved accuracies of image alignment and subsequent resolution of the final reconstruction (Extended Data Table 1).

A model of TmrAB was constructed based on homology modelling and molecular dynamics flexible fitting. The crystal structure of a heterodimeric ABC exporter from a thermophilic bacterium, *Thermotoga maritima*, TM287/288 (ref. 8), was used as the starting template. TM287 shares 31.3% sequence identity with TmrB, whereas TM288 shares 37.8% sequence identity with the TmrA. After consideration of two-dimensional class averages (Fig. 1c), immunoblotting data (Fig. 1d) and three-dimensional reconstruction of TmrAB–BA6 (Extended Data Fig. 4), we concluded that BA6 binds to the tip of the NBD of TmrB, opposite the AH5 binding site. This allowed us to position unambiguously the homology model into the density map as well as to determine the handedness of the three-dimensional reconstruction. The model was fitted into the TmrAB–AH5 density map by molecular dynamics flexible fitting<sup>19</sup>, yielding the final model of TmrAB (Extended Data Fig. 7). In both the three-dimensional density map and the fitted model, a cavity is evident that is accessible to the cytoplasm, suggesting that TmrAB is in an inward-facing conformation (Fig. 3a). The model also confirms that AH5 and BA6 interact with the NBD of TmrA and TmrB respectively. Such Fab binding inhibits ATPase activity of TmrAB (Extended Data Fig. 8), thus locking TmrAB in the current conformation.

In the three-dimensional reconstructions, the detergent micelle forms a torus in which two helices lie in the surface, resulting in an unusual bi-lobed micelle with two halves tilted from each other by about 30° (Fig. 2a). This separation is mediated in both monomers by TM4, which curves outwards with TM5 and protrudes closer to the head groups of  $\alpha$ -DDM (Fig. 2c). The TMDs contain two subdivisions, each of which is composed of TM1, TM2, TM3 and TM6 of one monomer and TM4 and TM5 of the other. The splitting of the micelle matches the subdivisions of the TMDs in this conformation.

A large cavity is located on one side of the TMDs, surrounded by TM1–3 and TM6 of TmrA and TM4–6 of TmrB (Fig. 3a). Significant density (visualized at 5 $\sigma$ ) was observed in the cavity bound to TM5 of TmrB (Fig. 3b and Extended Fig. 3a), analogous to the substrate binding sites identified for glutathione ATM1-type transporters<sup>20,21</sup>. This density has sufficient size to accommodate molecules as large as a Hoechst 33342 molecule, a known substrate for TmrAB<sup>5</sup>. While the identity of the molecule that contributed to this density is unknown, we speculate that it is attributable to an unknown molecule co-purified with TmrAB, molecules of DDM detergent or lipids, suggesting that the cavity contains the cytoplasmic substrate-binding site. Although the cavity is open to the cytosol, it also has a small, lateral V-shaped gap (formed by TM4 and TM6 of TmrB) that provides an additional entry point from the inner leaflet of the surrounding membrane. However, this gap is considerably smaller than the opening found in a wide-open, V-shaped apo conformation, such as in the open apo state of MsbA<sup>22</sup>.

The structure of TmrAB reported here is in an inward-facing state in which two NBDs are in contact with each other via their carboxy (C)-terminal helices. This conformation is similar to the substrate-bound nucleotide-free state of ATM1-type exporters<sup>20,21</sup> or a eukaryotic



**Figure 3 | Atomic model of TmrAB showing the laterally open inward-facing conformation.** **a**, Representation of internal volume and opening to the external surface of the transporter. TmrA is coloured salmon and TmrB is coloured cyan. **b**, Two different views (tilted around the axis perpendicular to the membrane plane) of the substrate-binding cavity in the TMDs. Density bound to helix H4 (Tyr 187) and H5 (His 246) of TmrB was observed at a threshold of 5 $\sigma$ . It has the size to accommodate a Hoechst 33342 molecule, which is a known cargo molecule of TmrAB, but was not added during the protein purification. The cargo-like density is inside the cavity but near the inner leaflet of the membrane. This position suggests a possible substrate pathway. **c**, Ribbon diagram of the TmrAB NBDs in the nucleotide-free state. Predicted locations of two pairs of cysteine mutations are marked. **d**, Non-reducing SDS–polyacrylamide gel electrophoresis (SDS–PAGE) gel demonstrating disulphide cross-linking of both double cysteine mutants in the apo state, and showing a clear difference in cross-linking behaviour between nucleotide-free and -bound TmrAB.

P-glycoprotein homologue<sup>23</sup>. We confirmed that the isolated TmrAB truly is nucleotide free (Extended Data Fig. 8). Therefore, in the nucleotide-free TmrAB structure, two NBDs remain in contact with each other via the C-terminal helices of both NBDs (Fig. 2d), burying a total surface area of ~980 Å<sup>2</sup> (see Methods) with a minimal distance between the C $\alpha$  positions of the two C-terminal helices of ~5 Å. Disulphide bond cross-linking of nucleotide-free TmrAB with two independent cysteine mutation pairs in their C-terminal helices (TmrA–L585C/TmrB–G564C, or TmrA–A591C/TmrB–A567C) validated this interaction (Fig. 3c, d and Extended Data Fig. 9). Consequentially, the two subdivisions of the TMDs are not widely separated. Indeed, no TmrAB particles were observed in a wide-open, V-shaped conformation as found in some crystal structures of nucleotide-free ABC exporters where the two NBDs are separated by a distance of 20–40 Å (refs 9, 22, 24, 25).

Connecting conformational states of a given ABC transporter to its substrate transport cycle requires structures at all distinct functional

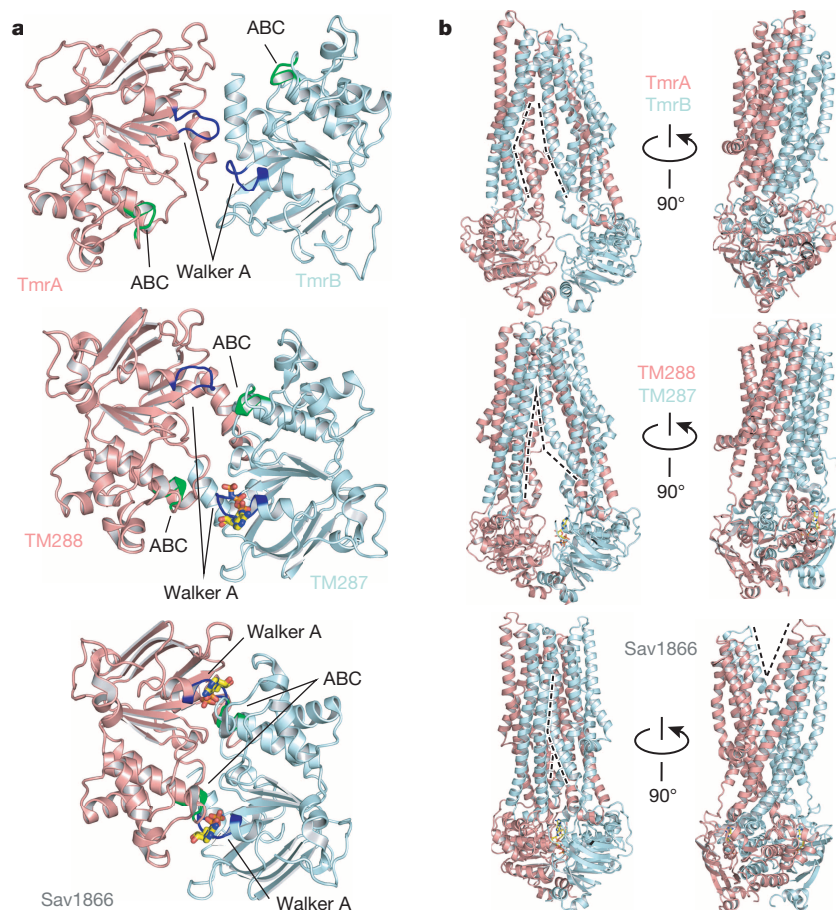


states, which are currently not available for most transporters. Therefore we were limited to comparing the structures from different members of the family. The well-resolved helices and distinct features of our reconstruction ensure that the structure of TmrAB reported here is of sufficient precision to compare on a secondary structural level with structures of other ABC exporters, the AMPPNP-bound intermediate inward-facing structure of TM287/288 (ref. 8) and the ADP-bound outward-facing structure of homodimeric Sav1866 (ref. 26), determined by crystallography. Such comparison suggests a trajectory of conformational changes connecting these states (Supplementary Videos 1 and 2). In TmrAB, the Walker A motifs in the two NBDs face each other but do not face the ABC signature motifs of the opposite NBDs (Fig. 4a top). In AMPPNP-TM287/288, the NBDs are in a different dimer contact state with the Walker A and ABC signature motifs of the opposite NBDs facing each other (Fig. 4a middle). In the outward-facing Sav1866 structure, the NBDs are in a further closed dimer state (Fig. 4a bottom) to sandwich two nucleotides between the NBDs as if in a hydrolytically competent-like state. This suggests that the NBDs may slide and rotate against each other as they bind, then sandwich ATP molecules between the domains during the transition from the inward- to outward-facing conformation in a mechanism similar to that of the ABC importers<sup>27</sup>.

The relative NBD positions are associated with reorganizations of the TMDs, especially TM4 and TM6. In TmrAB, TM4 and TM6 of chain B show the furthest separation of  $\sim 18$  Å (Fig. 4b top), which corresponds

to the lateral cavity opening shown in the open volume calculation of Fig. 3 as well as the substrate position. The lateral cavity opening is present in the intermediate state structure of TM287/288 but has narrowed to a maximal separation of 10 Å (Fig. 4b middle) between the hinges of TM4 and TM6. In the outward-open conformation of Sav1866, the cavity is completely occupied by TM4 and TM6, leaving an opening exposed to the extracellular surface (Fig. 4b bottom). The rearrangements in the transmembrane domain, specifically TM4 and TM6, work in concert to close the lateral membrane gate and open a perpendicular extracellular gate. These movements are coupled to sliding and rotation of the NBDs from an orientation with Walker A motifs in contact with each other to an orientation wherein the Walker A motifs contact the opposing ABC motif (Supplementary Videos 3 and 4).

Here, we have used single-particle cryo-EM to determine a subnanometre-resolution structure of detergent-solubilized TmrAB in the nucleotide-free state without influence from crystal packing. The resulting structure revealed an asymmetric inward-facing state with NBDs in contact but not in a position for ATP hydrolysis. Compared with another membrane protein recently determined to 3.4 Å resolution (TRPV1,  $\sim 300$  kDa and C4 symmetry)<sup>11</sup>, TmrAB is a more challenging target because of its smaller size ( $\sim 135$  kDa) and its pseudo-symmetry. The structure reported here greatly extends the capability of single-particle cryo-EM. At this resolution all helices in TMDs and NBDs are unambiguously resolved, affording the determination of a reliable model based



**Figure 4 | Comparison of the inward-facing conformation of TmrAB with the intermediate inward-facing structure of TM287/288 and the outward-facing structure of Sav1866. a,** Conformational changes in the NBDs viewed from the membrane towards the cytosol. The Walker A motifs (blue) do not face the ABC signature motifs (green) in the atomic model of TmrAB (top) while they do in the TM287/288 structure (middle) where a single AMPPNP molecule (shown as sticks) is bound predominantly in one NBD.

In the Sav1866 structure (bottom), the two motifs from the opposite domains come close to sandwich two ADP molecules (shown as sticks). **b,** Changes in lateral opening to the cavity. The lateral opening between TM4 and TM6 shown by dotted lines (left) narrows down from the TmrAB model (top) to the TM287/288 structure (middle), and completely closes in the Sav1866 structure (bottom), resulting in an opening to the extracellular space.

on the well-established workflow of homology modelling and flexible docking<sup>19</sup>. It was made possible by taking advantage of three novel technologies: (1) Fab-assisted single-particle cryo-EM<sup>13</sup>, (2) newly developed cryo-EM methodologies including a direct electron-detection camera and (3) maximum likelihood based classification and refinement<sup>28</sup>. Each of these technological advancements improves accuracy of image alignments and the resolutions of the final reconstructions (Extended Data Table 1). In theory, it is possible to achieve  $\sim 3$  Å resolution structure of proteins as small as  $\sim 100$  kDa (ref. 12). In practice, there are still major technological limitations that prevent achieving this goal, even with recent advances. Nevertheless, we expect that the combined approaches used in this study will be a powerful strategy for capturing the molecular features of membrane proteins in discrete functional states that are difficult to capture by crystallization. These structural snapshots will aid in the elucidation of membrane protein mechanisms while advancing the limits in structural determination by single-particle cryo-EM.

While this paper was under review, a nucleotide-free structure of TM287/288 was reported<sup>29</sup>, which is almost identical to the singly AMPPNP-bound structure (r.m.s.d. of 0.636 Å). It supports our conclusion that the NBDs remain in contact in the nucleotide-free state, but cross-linking behaviour (Fig. 3c) suggests a noticeable conformational change in TmrAB upon AMPPNP binding. Fully addressing such differences requires structures of the same protein at all functional states, which single-particle cryo-EM holds the promise to deliver.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 27 March; accepted 17 September 2014.**

**Published online 2 November 2014.**

- Rees, D. C., Johnson, E. & Lewinson, O. ABC transporters: the power to change. *Nature Rev. Mol. Cell Biol.* **10**, 218–227 (2009).
- Schmitt, L. & Tampé, R. Structure and mechanism of ABC transporters. *Curr. Opin. Struct. Biol.* **12**, 754–760 (2002).
- Gottesman, M. M. & Ambudkar, S. V. Overview: ABC transporters and human disease. *J. Bioenerg. Biomembr.* **33**, 453–458 (2001).
- Parcej, D. & Tampé, R. ABC proteins in antigen translocation and viral inhibition. *Nature Chem. Biol.* **6**, 572–580 (2010).
- Zutz, A. et al. Asymmetric ATP hydrolysis cycle of the heterodimeric multidrug ABC transport complex TmrAB from *Thermus thermophilus*. *J. Biol. Chem.* **286**, 7104–7115 (2010).
- Higgins, C. F. & Gottesman, M. M. Is the multidrug transporter a flippase? *Trends Biochem. Sci.* **17**, 18–21 (1992).
- vanHelvoort, A. et al. MDR1 P-glycoprotein is a lipid translocase of broad specificity, while MDR3 P-glycoprotein specifically translocates phosphatidylcholine. *Cell* **87**, 507–517 (1996).
- Hohl, M., Briand, C., Grutter, M. G. & Seeger, M. A. Crystal structure of a heterodimeric ABC transporter in its inward-facing conformation. *Nature Struct. Mol. Biol.* **19**, 395–402 (2012).
- Shintre, C. A. et al. Structures of ABCB10, a human ATP-binding cassette transporter in apo- and nucleotide-bound states. *Proc. Natl Acad. Sci. USA* **110**, 9710–9715 (2013).
- Ward, A. B. et al. Structures of P-glycoprotein reveal its conformational flexibility and an epitope on the nucleotide-binding domain. *Proc. Natl Acad. Sci. USA* **110**, 13386–13391 (2013).
- Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107–112 (2013).
- Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**, 171–193 (1995).
- Wu, S. et al. Fabs enable single particle cryoEM studies of small proteins. *Structure* **20**, 582–592 (2012).
- Paduch, M. et al. Generating conformation-specific synthetic antibodies to trap proteins in selected functional states. *Methods* **60**, 3–14 (2013).
- Kim, J., Stroud, R. M. & Craik, C. S. Rapid identification of recombinant Fabs that bind to membrane proteins. *Methods* **55**, 303–309 (2011).
- Li, X. et al. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**, 584–590 (2013).
- Scheres, S. H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nature Methods* **9**, 853–854 (2012).
- Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11**, 63–65 (2014).
- Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673–683 (2008).
- Lee, J. Y., Yang, J. G., Zhitnitsky, D., Lewinson, O. & Rees, D. C. Structural basis for heavy metal detoxification by an Atm1-type ABC exporter. *Science* **343**, 1133–1136 (2014).
- Srinivasan, V., Pierik, A. J. & Lill, R. Crystal structures of nucleotide-free and glutathione-bound mitochondrial ABC transporter Atm1. *Science* **343**, 1137–1140 (2014).
- Ward, A., Reyes, C. L., Yu, J., Roth, C. B. & Chang, G. Flexibility in the ABC transporter MsbA: alternating access with a twist. *Proc. Natl Acad. Sci. USA* **104**, 19005–19010 (2007).
- Kodan, A. et al. Structural basis for gating mechanisms of a eukaryotic P-glycoprotein homolog. *Proc. Natl Acad. Sci. USA* **111**, 4049–4054 (2014).
- Aller, S. G. et al. Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science* **323**, 1718–1722 (2009).
- Jin, M. S., Oldham, M. L., Zhang, Q. & Chen, J. Crystal structure of the multidrug transporter P-glycoprotein from *Caenorhabditis elegans*. *Nature* **490**, 566–569 (2012).
- Dawson, R. J. & Locher, K. P. Structure of a bacterial multidrug ABC transporter. *Nature* **443**, 180–185 (2006).
- Pinkett, H. W., Lee, A. T., Lum, P., Locher, K. P. & Rees, D. C. An inward-facing conformation of a putative metal-chelate-type ABC transporter. *Science* **315**, 373–377 (2007).
- Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
- Hohl, M. et al. Structural basis for allosteric cross-talk between the asymmetric nucleotide binding sites of a heterodimeric ABC exporter. *Proc. Natl Acad. Sci. USA* **111**, 11025–11030 (2014).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was supported by grants from the National Institutes of Health (R01GM098672, S10RR026814 and P50GM082250 to Y.C., 1P41CA196276-01 to C.S.C., P50GM073210 to R.M.S. and C.S.C., and R37GM024485 to R.M.S.), the University of California San Francisco Program for Breakthrough Biomedical Research (to Y.C.), and the German Research Foundation (SFB 807, SFB 902 and TA157/7 to R.T.) as well as the European Drug Initiative on Channels and Transporters (EDICT to R.T.) funded by the European Commission Seventh Framework Program.

**Author Contributions** J.K. identified, expressed, purified and characterized all Fabs used in this study, and generated TmrAB–Fab complexes. S.W. performed all cryo-EM experiments, including data acquisition and processing. T.M.T. and C.M. expressed and purified TmrAB, and purified TmrAB–Fab complexes. T.M.T. performed cross-linking experiments. C.M. expressed and purified TmrAB for the generation and initial screening of all Fabs. S.B.S. performed initial characterization of all Fabs. M.B.W. performed high-performance liquid chromatography (HPLC) experiments. Y.R.-C. performed mutagenesis experiments. J.K., S.W., T.M.T. and Y.C. analysed data. J.K., S.W., T.M.T., M.B.W., R.M.S., R.T., C.S.C. and Y.C. participated in discussion and wrote the manuscript.

**Author Information** All three-dimensional cryo-EM density maps have been deposited in the Electron Microscopy Data Bank under accession numbers EMD-6085 (TmrAB–AH5), EMD-6086 (TmrAB–BA6) and EMD-6087 (TmrAB). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.M.S. ([stroud@msg.ucsf.edu](mailto:stroud@msg.ucsf.edu)), R.T. ([tampe@em.uni-frankfurt.de](mailto:tampe@em.uni-frankfurt.de)), C.S.C. ([Charles.Craik@ucsf.edu](mailto:Charles.Craik@ucsf.edu)) or Y.C. ([yicheng@ucsf.edu](mailto:yicheng@ucsf.edu)).

## METHODS

**Sample preparation.** TmrAB was expressed and purified similarly as described previously<sup>5</sup>. Protein was extracted by using 1%  $\beta$ -DDM in a buffer containing 50 mM Tris pH 8.0, 300 mM NaCl, and 1 mg ml<sup>-1</sup> iodoacetamide was added followed by purification on Ni<sup>2+</sup> resin in 50 mM Tris pH 8.0, 300 mM NaCl, 0.05%  $\alpha$ -DDM and finished with size-exclusion chromatography on a Superdex S200 (GE Healthcare) with 20 mM HEPES pH 7.0, 150 mM NaCl 0.05%  $\alpha$ -DDM. For ELISA, purified TmrAB was biotinylated, using EZ-Link NHS-Chromogenic-Biotin (Pierce) as described previously<sup>15</sup>.

Phagemids were used for myc-tagged Fab (Fab-myc) expression without further modification<sup>15</sup>. For constructs to express Flag-tagged Fab (Fab-Flag), the myc tag in the phagemid was replaced with the Flag tag by PCR site-directed mutagenesis with a primer encoding DYKDDDDK. The amplified PCR products were cloned in at BstEII and NotI sites in the phagemids.

Fabs were expressed in *Escherichia coli* with IPTG induction. Phagemids encoding each Fab gene were transformed into BL21 (DE3) Gold (Stratagene). Transformed cells were grown in 2YT containing 100  $\mu$ g ml<sup>-1</sup> ampicillin and 2% glucose (2YT Amp Glu) at 250 r.p.m. and 30 °C overnight. The overnight culture was aliquoted into fresh 2YT Amp 0.1% Glu such that attenuation ( $D_{600\text{ nm}}$ ) was around 0.05. The cultures were grown at 250 r.p.m. and 37 °C until the log phase where  $D_{600\text{ nm}}$  was between 0.5 and 0.8 and induced with 1 mM IPTG. The induced cultures were grown at 200 r.p.m. and 20 °C overnight.

*E. coli*-expressed Fabs were subject to periplasmic protein fractionation by osmotic shock. Cell pellets of the induced cultures were re-suspended in ice-cold TES buffer (0.2 M Tris pH8, 0.5 mM EDTA, 0.5 M sucrose), followed by an equal volume of ice-cold double-distilled water containing protease inhibitor cocktail (complete EDTA-free, Roche). The cell re-suspension was incubated on ice for 30 min with gentle swirling every 10 min and spun down at 13,000g for 15 min. The supernatant was collected for further purification.

The periplasmic fraction was subject to affinity chromatography by a batch procedure using Ni<sup>2+</sup>-NTA agarose resin (Qiagen) and a standard protocol recommended by the manufacturer. Purified Fabs were dialysed against PBS buffer. For complex formation, the Fab samples were further purified by size-exclusion chromatography using Superdex 75 (Pharmacia Biotech). For cryo-EM, TmrAB and Fab were mixed in a 1:2 molar ratio respectively, incubated for 30 min at room temperature (~22 °C) and subject to size-exclusion chromatography, using Superdex 200 (Pharmacia Biotech). All samples for cryo-EM were prepared and stored in 20 mM HEPES buffer, pH 7 containing 150 mM NaCl and 0.05%  $\alpha$ -DDM.

**Fab selection.** Our Fab selection procedure has been described previously<sup>15</sup>. For a brief summary, the procedure included indirect immobilization of biotinylated TmrAB to streptavidin-coated magnetic beads, three rounds of selection of phage displayed Fabs that bind to TmrAB, and identification of binders by an ELISA screen, which used Fabs with unknown concentrations. In the Fab nomenclature, the first letters designated plates, and the second letters and the numbers designated well positions.

**Qualitative and competitive ELISA.** Qualitative ELISA was done as described previously<sup>15</sup>. TmrAB was coated in the wells of Maxisorp plates (Nunc) and Fab solutions at unknown concentrations were added to the wells for binding. Fab solutions were prepared by serial dilutions of unpurified Fabs to yield series of relative concentrations.

Competitive ELISA was done in the same way as qualitative ELISA analysis using relative Fab concentrations<sup>15</sup>, with the following modifications. (1) Equal volumes of Fab-myc and Fab-Flag solutions were added for each binding reaction. Unpurified Fab-myc at varying concentrations was prepared by serial dilution. Unpurified undiluted Fab-Flag was kept at a constant concentration. (2) Duplicates for each binding reaction were set up. (3) Anti-c-myc-peroxidase (Roche) was added in one reaction, and anti-Flag M2-peroxidase (Sigma) was added in the other. In Fig. 1, binding reactions for each graph contain AD12-myc at varying concentrations and Fab-Flag at constant concentrations, and TmrAB immobilized on wells. AD12-myc concentration in each binding reaction was relative to undiluted supernatant of AD12-myc culture that was considered 1. Normalized ELISA signals = (ELISA signal of Fab-Flag in each binding reaction)/(ELISA signal of Fab-Flag when [AD12-myc] = 0); \*Normalized maximum binding of Fab-Flag = 1. Bound Fabs were detected by anti-Flag M2-peroxidase.

**Analytical HPLC measurements.** TmrAB ATPase activity and nucleotide stoichiometry measurements were performed with reversed-phase HPLC using an Agilent 1100 system. For all measurements, TmrAB and nucleotide (ATP/ADP) standards underwent organic extraction in the same manner with 25:24:1 (v/v) phenol:chloroform:isoamyl alcohol. The aqueous layer was resolved using a C<sub>18</sub> column (Vydac, 218TP104) and gradient elution (solvent A: 100 mM KH<sub>2</sub>PO<sub>4</sub>, pH 6.0; solvent B: 90% methanol) with absorbance measurements recorded at 254 nm.

**Cysteine cross-linking.** TmrAB double cysteine mutants were created to test the proximity of the NBDs in the apo state. Two double-cysteine mutant constructs were

generated, TmrA\_A591C/TmrB\_A567C and TmrA\_L595A/TmrB\_G564C. A591C/A567C was designed using the primers 5'-GGCCAAGGGCGGCTACTACGCC TGCTTGATACCGGCTCCAGTTCAG-3' and 5'-CAGGCCGAGGCGCTCTA CTGCGAGATGGACCGCCTGCAG-3', and L565C/G564C using 5'-CCGCCT TGTACCGGTGCCAGTTCAGGAGGC-3' and 5'-GAGAGCCTCCTTCAGG CCGGATGCGCTCTACGCGGAGATGGACCGCCTG-3'. Note that TmrAB contains an exposed native cysteine residue (TmrA-416C) that could not be removed.

The double cysteine mutants were purified without nucleotides following the similar procedure used for the wild-type TmrAB, with  $\beta$ -DDM (~1–1.05%) used for protein extraction and no iodoacetamide added. Cross-linking of TmrAB with double cysteine mutations was assessed in the presence or absence of 1 mM Cu<sup>2+</sup> phenanthroline, 0.7 mM AMPNP or ATP/Na<sup>+</sup> orthovanadate (5 mM) with or without 5 mM MgCl<sub>2</sub> as indicated in the label of Extended Data Fig. 9b. All nucleotides were heated to 68 °C, the permissive temperature for maximal ATP hydrolysis, for 3.5 min followed by immediate cooling on ice then 1 mM copper phenanthroline (Sigma) -induced cross-linking for 30 min on ice. Cross-linking was quenched with 125 mM EDTA in a non-reducing SDS loading buffer and immediately loaded onto SDS-PAGE for analysis.

The expression level of L595C/G564C was much lower than that of A591C/A567C. Nevertheless, L595C/G564C could be purified and was assessed for cross-linking with and without the oxidizing reagent Cu<sup>2+</sup> phenanthroline at a lower concentration (567 nM) because of lower yields. Cross-linking in both cases proceeded to >99% as judged by SDS-PAGE.

The A591C/A567C mutant activity was compared with the wild-type protein following cross-linking with copper phenanthroline followed by buffer exchange on Millipore spin concentrators. Samples were incubated with 73.28 nM TmrAB, 2.09 mM MgCl<sub>2</sub> and 1 mM copper phenanthroline for 30 min, followed by three ~100-fold dilutions with the purification buffer (20 mM HEPES, 150 mM NaCl, 0.05%  $\alpha$ -DDM). Samples were then incubated with ATP at a final concentration of 70 nM transporter, 2 mM MgCl<sub>2</sub>, and 250  $\mu$ M ATP (pH 7.0) at 60 °C for 1 h in preparation for analytical HPLC analysis.

**Negative-stain electron microscopy.** Each Fab and TmrAB was mixed in a 2:1 molar ratio in pH 7.4 PBS containing 0.05%  $\beta$ -DDM, incubated at room temperature (~22 °C) for 5 min, and negatively stained using uranyl formate. Electron microscope grids of negatively stained samples were prepared as previously described<sup>30</sup>. Specifically, 2.5  $\mu$ l samples of detergent-solubilized TmrAB alone or in complex with Fabs were applied to glow-discharged electron microscope Cu grids covered by a thin layer of continuous carbon film and were stained with 0.75% (w/v) uranyl formate. Negatively stained electron microscope grids were observed on a Tecnai T12 microscope (FEI) operated at 120 kV. Images were recorded at a nominal magnification of  $\times 52,000$  using a 4K  $\times$  4K CCD (charge-coupled device) camera (UltraScan 4000, Gatan), corresponding to a pixel size of 2.21 Å on the specimen. All images were binned to the final pixel size of 4.42 Å for further processing. Particles were selected manually and were subjected to six cycles of multi-reference alignment and K-means classification using SPIDER<sup>31</sup>. Approximately 3,000 particles for the TmrAB-Fab complex were manually picked from each sample and used for two-dimensional analysis. Sixty classes were generated for TmrAB in complex with single Fabs except DH5, and all double Fabs. Fifty classes were generated for TmrAB in complex with DH5.

**Cryo-EM.** Cryo-EM grids were prepared by a standard plunge freezing procedure as described<sup>16</sup>. Specifically, Quantifoil grids (Quantifoil Micro Tools) were glow-discharged for 10 s and a 2  $\mu$ l sample was loaded onto the grid and plunge-frozen in liquid ethane cooled by liquid nitrogen using a VitroBot Mark III (FEI). CryoEM grids of the TmrAB-AH5, -BA6 complexes and TmrAB alone were imaged using a Tecnai TF20 electron microscope equipped with a field emission source (FEI) and operated at 200 kV. Images were collected at a nominal magnification of  $\times 80,000$  using a TemCam-F816 8k  $\times$  8k CMOS camera (TVIPS) corresponding to a pixel size of 0.935 Å on the specimen. All images were binned by a factor of 4 for further processing, resulting in a pixel size of 3.74 Å. Images were recorded with a defocus in the range from -2.0 to -3.5  $\mu$ m. Defocus values were determined for each micrograph using CTFFIND3 (ref. 32).

For the AH5 complex, an additional data set was collected using a Tecnai TF30 Polara equipped with a field emission source and operated at 300 kV. Images were recorded on a Gatan K2 Summit camera operated in super-resolution counting mode following the established dose fractionation data acquisition protocol<sup>16</sup>. Specifically, images from TF30 were recorded at a nominal magnification of  $\times 20,000$ , corresponding to a calibrated super resolution pixel size of 0.98 Å on the specimen. The dose rate on the camera was set to about eight counts (corresponding to ~9.9 electrons) per physical pixel per second. The total exposure time was 16 s, leading to a total accumulated dose of 41 electrons per square ångström on the specimen. Each image was fractionated into 32 subframes, each with an accumulation time of 0.5 s per frame. All dose-fractionated cryo-EM images were recorded using a semi-automated acquisition program, UCSFImage4 (written by X. Li). Images were recorded



with a defocus in a range from  $-2.5$  to  $-4.0$   $\mu\text{m}$ . All super-resolution counting images were binned by  $2 \times 2$  for further processing, resulting in a pixel size of  $1.96$   $\text{\AA}$ . All 32 motion-corrected subframes were averaged together to a single micrograph for subsequent processing. Defocus values were determined for each micrograph using CTFFIND3 (ref. 32). For the data set of the AH5 complex collected with TF20 microscope, the particles were manually picked. For all other data sets, we used a semi-automated particle-picking procedure similar to that previously described<sup>11,33</sup>. For each data set, we manually selected approximately 2,000 particles and generated two-dimensional class averages using reference-free classification followed by multiple rounds of multi-reference alignment and classification. Unique two-dimensional class averages were used as templates for automated particle picking. All picked particles were subject to visual inspection to remove all possible bad particles. Initial three-dimensional models were generated using the common lines method<sup>34</sup> and filtered to  $60$   $\text{\AA}$  resolution as the initial reference for further refinement. Three-dimensional reconstructions were calculated and refined using RELION<sup>28</sup>. The resolutions of final three-dimensional reconstructions were estimated using gold-standard Fourier shell correlation curve =  $0.143$  criterion<sup>17</sup>. The final map was sharpened by a resolution-dependent amplitude-scaling factor<sup>35</sup> and implemented in xMipp<sup>36</sup>. Local resolution was estimated using ResMap<sup>18</sup>. All cryo-EM image process parameters are listed in Extended Data Table 1, which also shows that adding Fab and using a direct electron-detection camera improved the alignment accuracy and resolution of three-dimensional reconstructions. Local correlations between maps were computed using the 'vop localCorrelation' function in UCSF Chimera (Extended Data Fig. 7). The difference in resolution may contribute to the low correlation at certain regions.

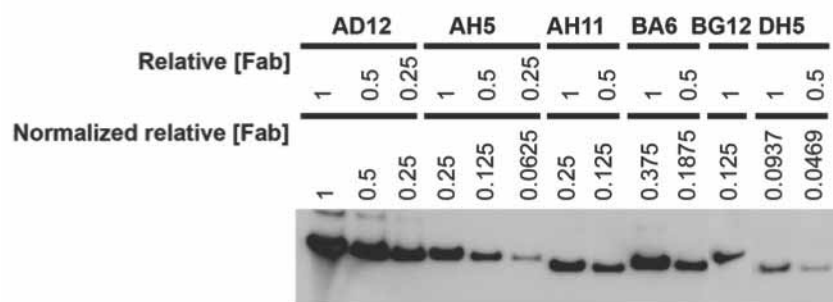
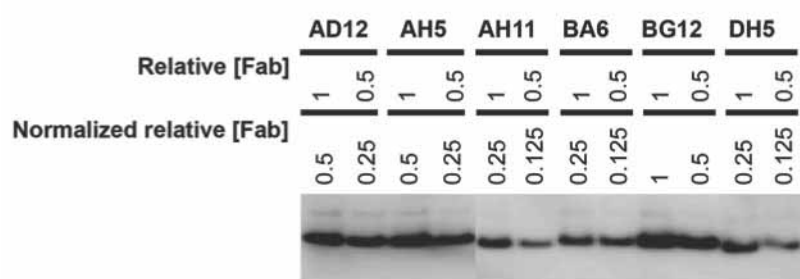
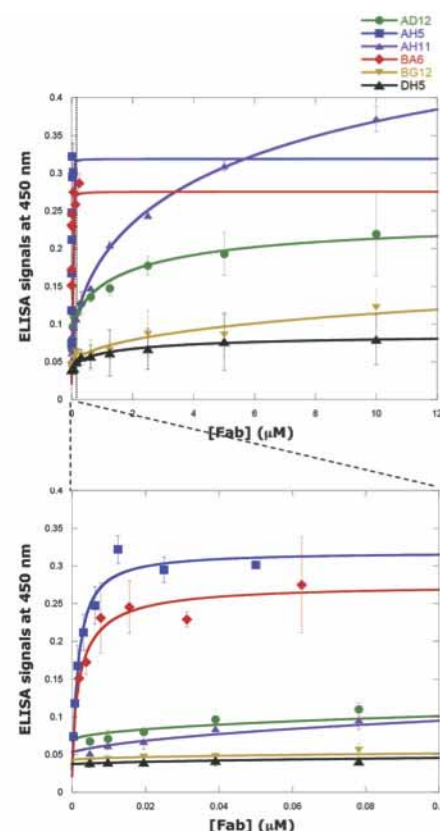
**Fitting of atomic structures.** A homology model of TmrAB was generated using SWISS-MODEL Automated Mode<sup>37</sup> and the atomic structure of heterodimeric ABC exporter TM287/TM288 with one AMPPNP bound (Protein Data Bank accession number 3QF4) used as the template. The atomic structure of a Fab (Protein Data Bank accession number 1M71) was used to fit the Fab density of either AH5 or BA6. Atomic models of TmrAB and Fab were docked into the three-dimensional reconstructions using the 'fit-in-map' function in UCSF Chimera<sup>38</sup>. Handedness was determined at this stage. The difference of fitting between the default map and its mirrored version is significant by visual inspection. The docked models were further refined by molecular dynamics flexible fitting<sup>19</sup>. The initial homology model shows an unstructured region between TM1 and TM2 of TmrA owing to a lack of corresponding sequence in the template. We analysed the sequence around the region using the Quick two-dimensional function in MPI Bioinformatics Toolkit<sup>39</sup>,

and part of it was predicted to form a helix by multiple prediction methods. Therefore, we built a short segment of helix for that part using the 'Build Structure' function in UCSF Chimera. That helical segment fitted well in the density map. All molecular graphics images were produced using UCSF Chimera<sup>38</sup>.

**Structure analysis.** Sequence identity between the TmrAB chains and the TM287/288 model were determined with the clustalW webserver. Internal cavities connected to the external environment were visualized using the 3V algorithm<sup>40</sup>. This procedure uses a  $1.5$   $\text{\AA}$  sphere to represent water and calculates a completely accessible volume. The protein surface environment is calculated with a  $12$   $\text{\AA}$  sphere that cannot access inner cavities, and the two volumes are subtracted to provide only those buried areas that water can enter from the external environment. In this calculation, only one large interconnected cavity was identified in the interior of TmrAB, connected to the surface via a cleft between TM4 and TM6. Buried surface-area calculations were calculated in Pymol.

30. Booth, D. S., Avila-Sakar, A. & Cheng, Y. Visualizing proteins and macromolecular complexes by negative stain EM: from grid preparation to image acquisition. *J. Vis. Exp.* <http://dx.doi.org/10.3791/3227> (22 December 2011).
31. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).
32. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).
33. Cao, E., Liao, M., Cheng, Y. & Julius, D. TRPV1 structures in distinct conformations reveal activation mechanisms. *Nature* **504**, 113–118 (2013).
34. Shaikh, T. R. *et al.* SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nature Protocols* **3**, 1941–1974 (2008).
35. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
36. Scheres, S. H., Nunez-Ramirez, R., Sorzano, C. O., Carazo, J. M. & Marabini, R. Image processing for electron microscopy single-particle analysis using XMIPP. *Nature Protocols* **3**, 977–990 (2008).
37. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195–201 (2006).
38. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
39. Biegert, A., Mayer, C., Remmert, M., Soding, J. & Lupas, A. N. The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res.* **34**, W335–W339 (2006).
40. Voss, N. R. & Gerstein, M. 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res.* **38**, W555–W562 (2010).



**a****Experiment A****Experiment B****b****c**

Fabs	Apparent $K_d$ ( $\mu$ M)	Relative affinity
AD12	2	++
AH5	0.001	+++++
AH11	9*	+
BA6	0.003	++++
BG12	n.d.	-
DH5	n.d.	-

**d**

	# complex particles	# total particles	% complex particles
TmrAB-AD12	1756	3279	53
TmrAB-AH5	2718	3171	86
TmrAB-AH11	1324	2986	44
TmrAB-BA6	2325	3009	77
TmrAB-BG12	576	2993	19
TmrAB-DH5	591	2677	22

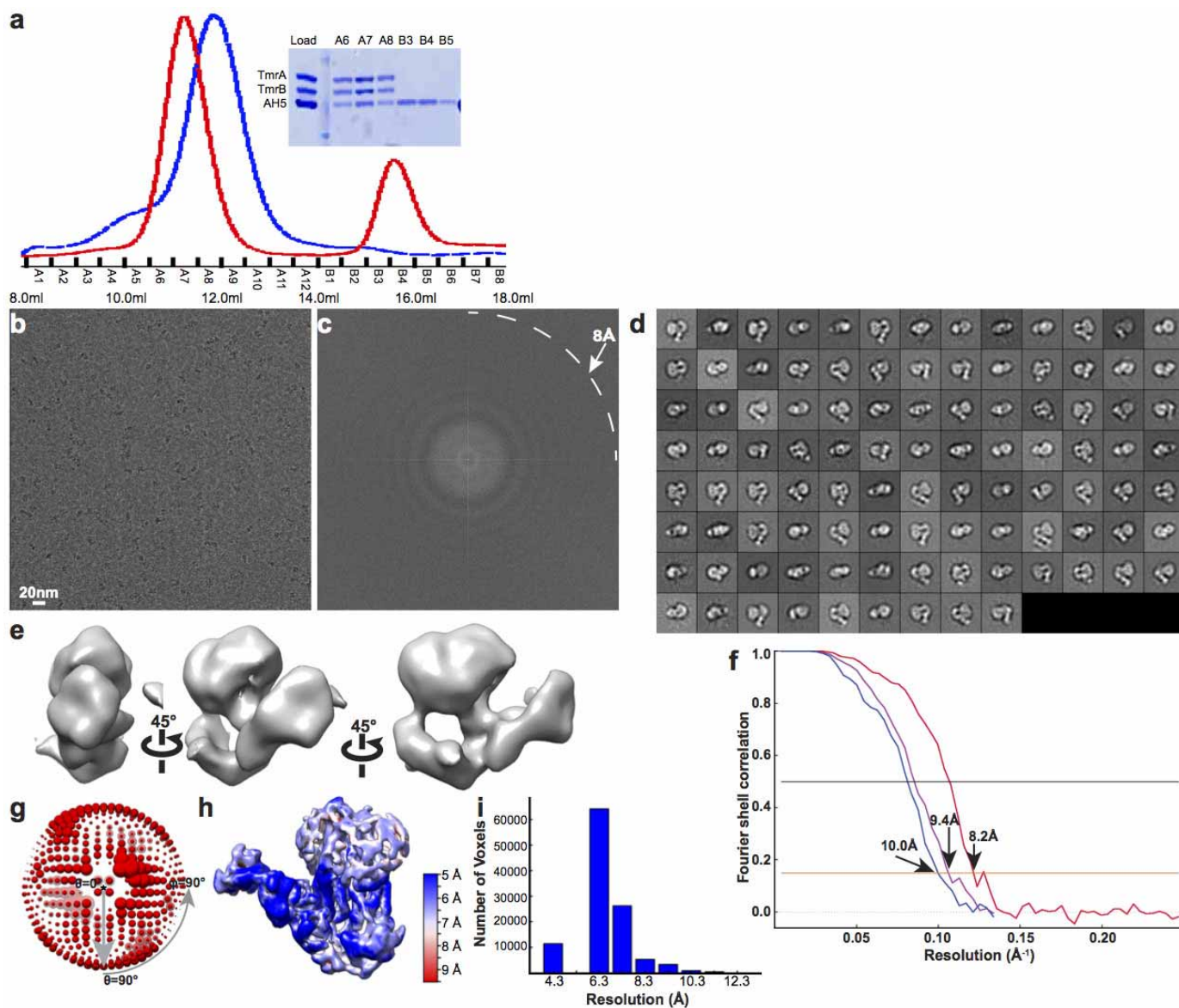
**Extended Data Figure 1 | Binding characterization of Fabs against TmrAB.**

**a**, Expression levels of Fabs used in ELISA in Fig. 1a. Expression levels were assessed by immunoblotting with anti-c-myc antibody and normalized against the highest expression level such that undiluted AD12 equals a normalized relative concentration of 1. Binding was monitored by anti-c-myc antibody.

**b**, ELISA with purified Fabs against TmrAB. Purified Flag-tagged Fabs (Fab-Flag) were used in binding reactions. Binding was monitored by anti-Flag M2-peroxidase. Experiments were repeated twice. **c**, Relative affinities of the Fabs. The ELISA signal data from Extended Data Fig. 1b were fitted to a bimolecular binding equation to produce binding curves and apparent dissociation constant ( $K_d$ ) values. AH5 showed the highest affinity, followed by BA6, AD12 and AH11. DH5 and BG12 did not show significant binding.

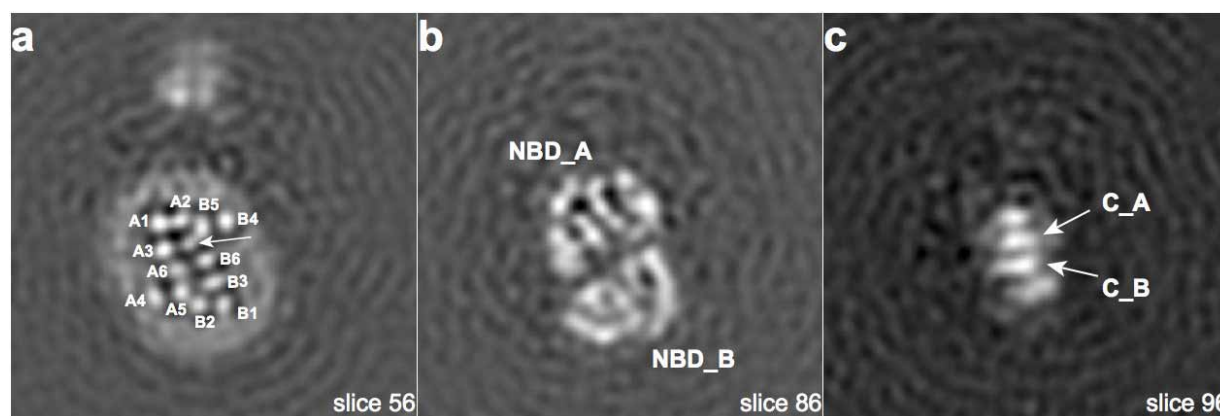
Apparent  $K_d$  values do not yield meaningful biophysical properties because the concentration of TmrAB participating in the binding reaction is unknown and

the transporter is not free in solution. However, comparison among the Fabs should be sufficient to determine their relative affinities. Unpurified DH5 showed significant binding (Extended Data Fig. 1a), whereas purified DH5 did not show significant binding; n.d., not determined. **d**, Negative-stain electron microscopy analysis of Fab + TmrAB mixes. Representative two-dimensional class averages include complex images that show clear Fab densities and images that do not. The two typical Fab views, the dumb-bell- and doughnut-shaped views, are indicated by yellow and red arrows respectively. Fab images indicate that Fabs are rigid and form rigid complexes. Percentages of complex particles were assessed by fractions of the numbers of images that clearly show Fab densities (# complex particles) to the total numbers of images that were included (# total particles) in two-dimensional class averages. The percentages correspond to relative affinity ranking determined by ELISA (Extended Data Fig. 1c).



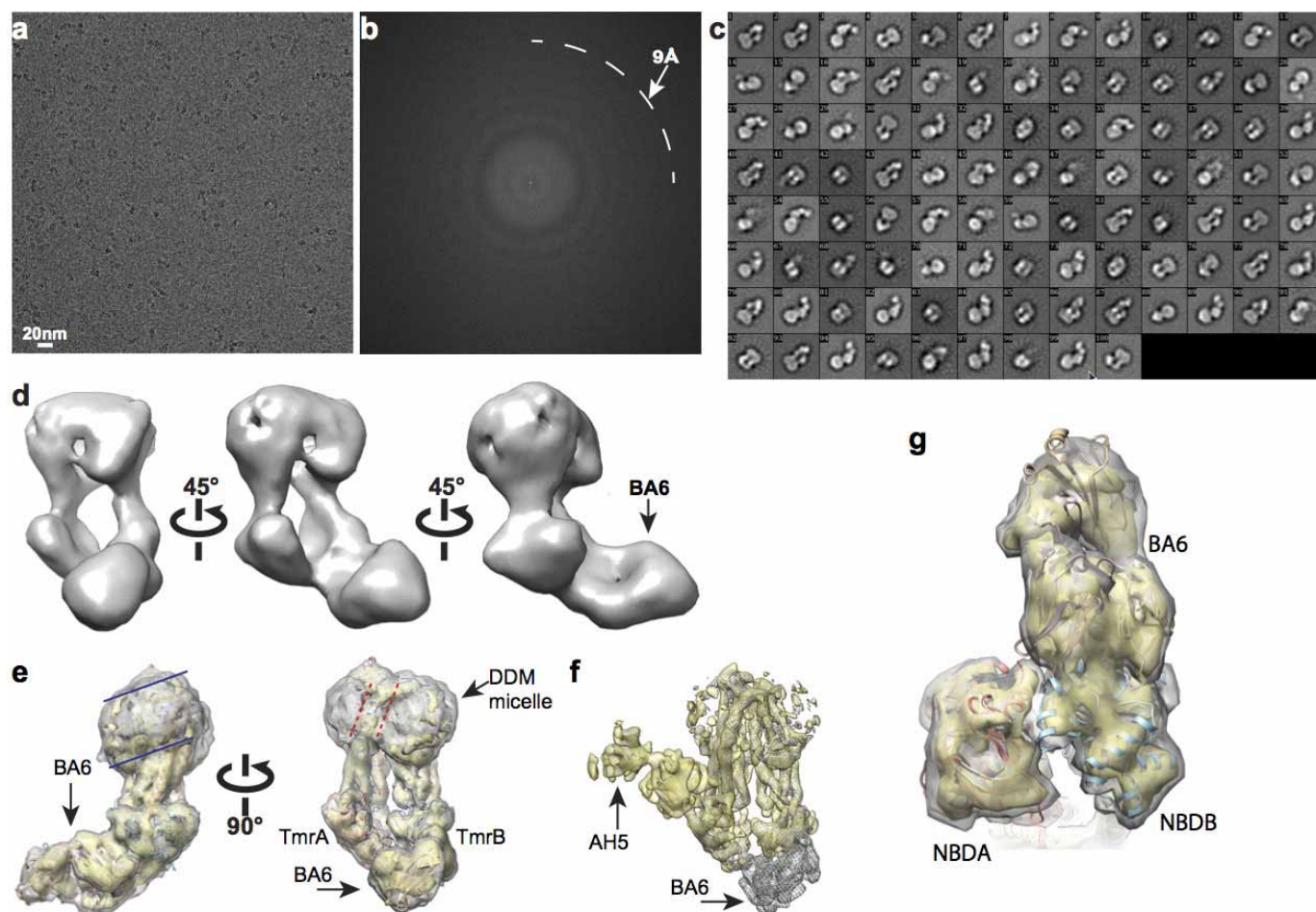
**Extended Data Figure 2 | Purification of  $\alpha$ -DDM-solubilized and single-particle cryo-EM of TmrAB-AH5 complex.** **a**, Elution profiles of TmrAB alone and TmrAB-AH5 from Superdex 200 are shown in blue and red curves respectively, showing a clear shift of the elution peak of the TmrAB-AH5 complex to a higher molecular mass position. The shifted peak corresponding to fractions A6–A8 contained TmrAB and AH5, confirmed by SDS-PAGE. Fractions B3–B5 correspond to unbound AH5 and the loading material was run for comparison. **b**, Raw micrograph of TmrAB-AH5 (~185 kDa) embedded in a thin layer of vitreous ice. **c**, Fourier power spectrum calculated from micrograph shown in **a**. **d**, Two-dimensional class averages of the

TmrAB-AH5 complex. Fab AH5 is clearly visible in many class averages. **e**, Initial three-dimensional reconstruction calculated from two-dimensional class averages using the common lines method implemented in SPIDER. **f**, Fourier shell correlation curves of TmrAB-AH5 (red), TmrAB-BA6 (purple) and TmrAB alone (blue). **g**, Euler angle distribution of all particles used in the final reconstruction. **h**, Final three-dimensional reconstruction coloured with local resolution. **i**, Voxel histogram corresponding to local resolution. The majority of voxels are at ~6–7 Å resolution. Estimation of local resolution that is too close to the Nyquist (3.9 Å) may not be accurate.



**Extended Data Figure 3 | Selected slice views of the three-dimensional reconstruction of TmrAB-AH5.** The views are oriented in parallel with the membrane plane. The numbers of slices are marked. **a**, All transmembrane

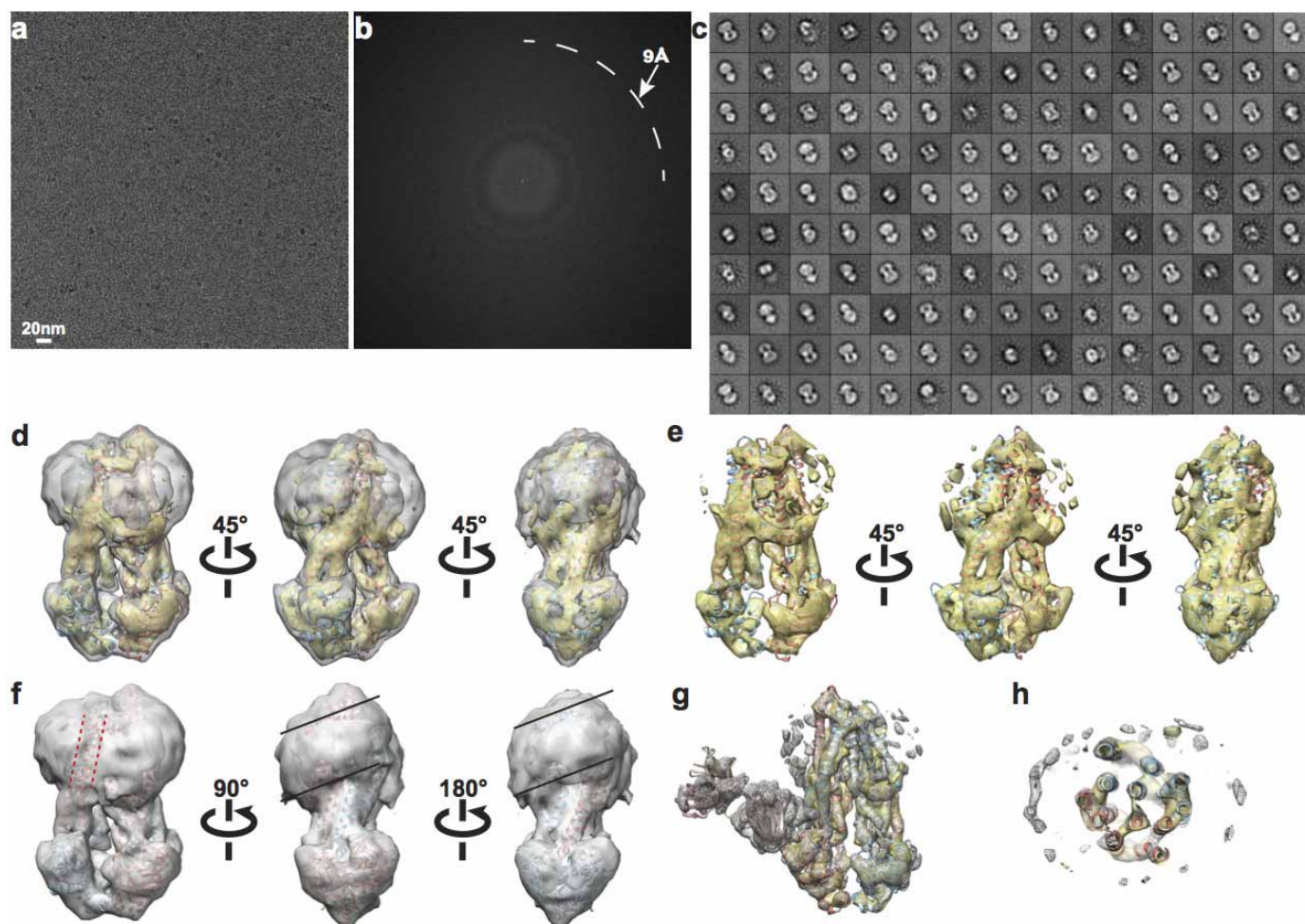
helices of both TmrA and TmrB are labelled. The arrow points to the extra density in the cavity. **b**, Two NBDs are in contact with each other. **c**, The C-terminal helices of TmrA and TmrB are in close contact.



**Extended Data Figure 4 | Single-particle cryo-EM of TmrAB-BA6 complex.** **a**, Raw micrograph of TmrAB-BA6 (~185 kDa) embedded in a thin layer of vitreous ice. Images were collected on a Tecnai TF20 microscope using a scintillator-based TVIPS  $8k \times 8k$  CMOS camera. **b**, Fourier power spectrum calculated from the micrograph shown in **a**. **c**, Two-dimensional class averages of the TmrAB-BA6 complex. Fab BA6 is clearly visible in many class averages. **d**, Initial three-dimensional reconstruction of TmrAB-BA6 determined using the common lines method implemented in SPIDER. **e**, Two different views of the final three-dimensional reconstruction of TmrAB-BA6

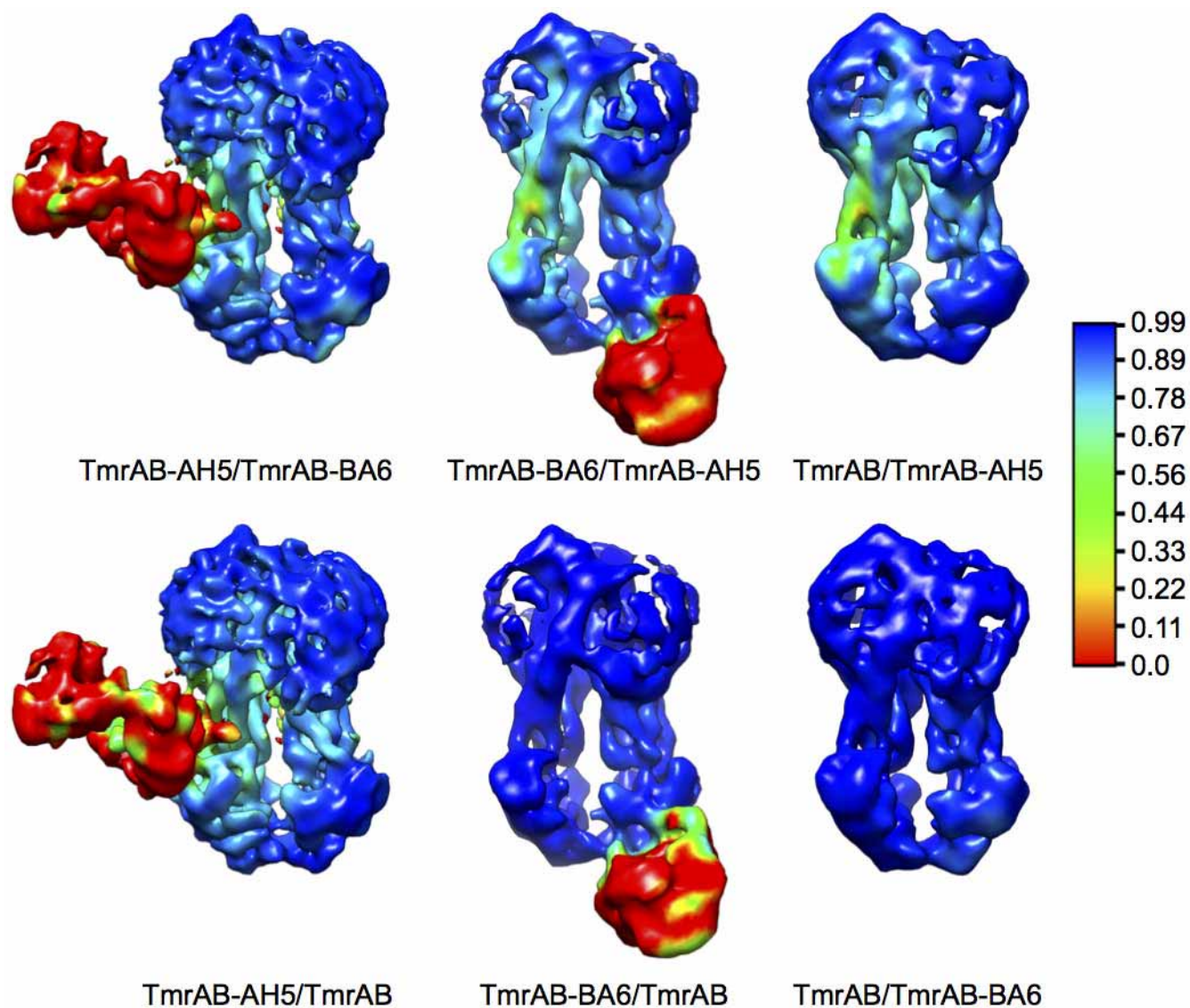
filtered to a resolution of  $9.4 \text{ \AA}$ . As in the three-dimensional reconstruction of TmrAB-AH5, the density of micelles is split into two halves and tilted with respect to each other. The orientation of micelle density is marked with a pair of black solid lines and the gap in the micelle density generated by the helix H4 from TmrB is marked with a pair of red dotted lines. **f**, Densities of TmrAB in the three-dimensional reconstructions of TmrAB-AH5 (khaki) and TmrAB-BA6 (grey mesh) overlap. Fabs AH5 and BA6 are indicated with arrows. **g**, An enlarged view to show the interface between TmrAB and BA6, which has a linear epitope in the NBD of TmrB.





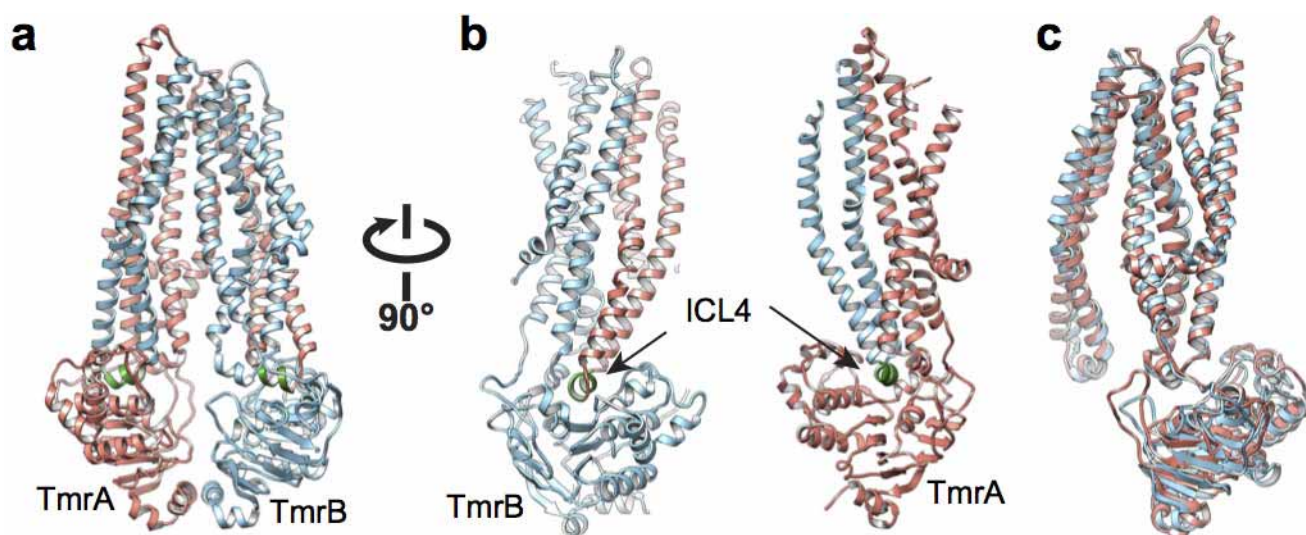
**Extended Data Figure 5 | Single-particle cryo-EM of TmrAB alone without Fab.** **a**, Raw micrograph of TmrAB alone ( $\sim 135$  kDa) embedded in a thin layer of vitreous ice. Images were collected on a Tecnai TF20 microscope using scintillator based TVIPS  $8k \times 8k$  CMOS camera. **b**, Fourier power spectrum calculated from micrograph shown in **a**. **c**, Two-dimensional class averages of TmrAB. **d–f**, Three different views of TmrAB three-dimensional reconstruction shown in different (low: grey; high: gold) contour levels.

Model of TmrAB (in ribbon diagram) was docked into the density map. The orientation of micelle density is indicated with pairs of solid black lines in **f** and the gap in the micelle is indicated with a pair of red dotted lines. **g, h**, Densities of TmrAB in the three-dimensional reconstructions of TmrAB alone (transparent khaki) and in complex with AH5 (grey mesh) overlap each other.



**Extended Data Figure 6 | Cross correlation between TmrAB-AH5, TmrAB-BA6 and TmrAB.** Left: density map of TmrAB-AH5 is coloured according to the value of local cross-correlation values of TmrAB-AH5 with TmrAB-BA6 (upper), with TmrAB (lower). Middle: density map of

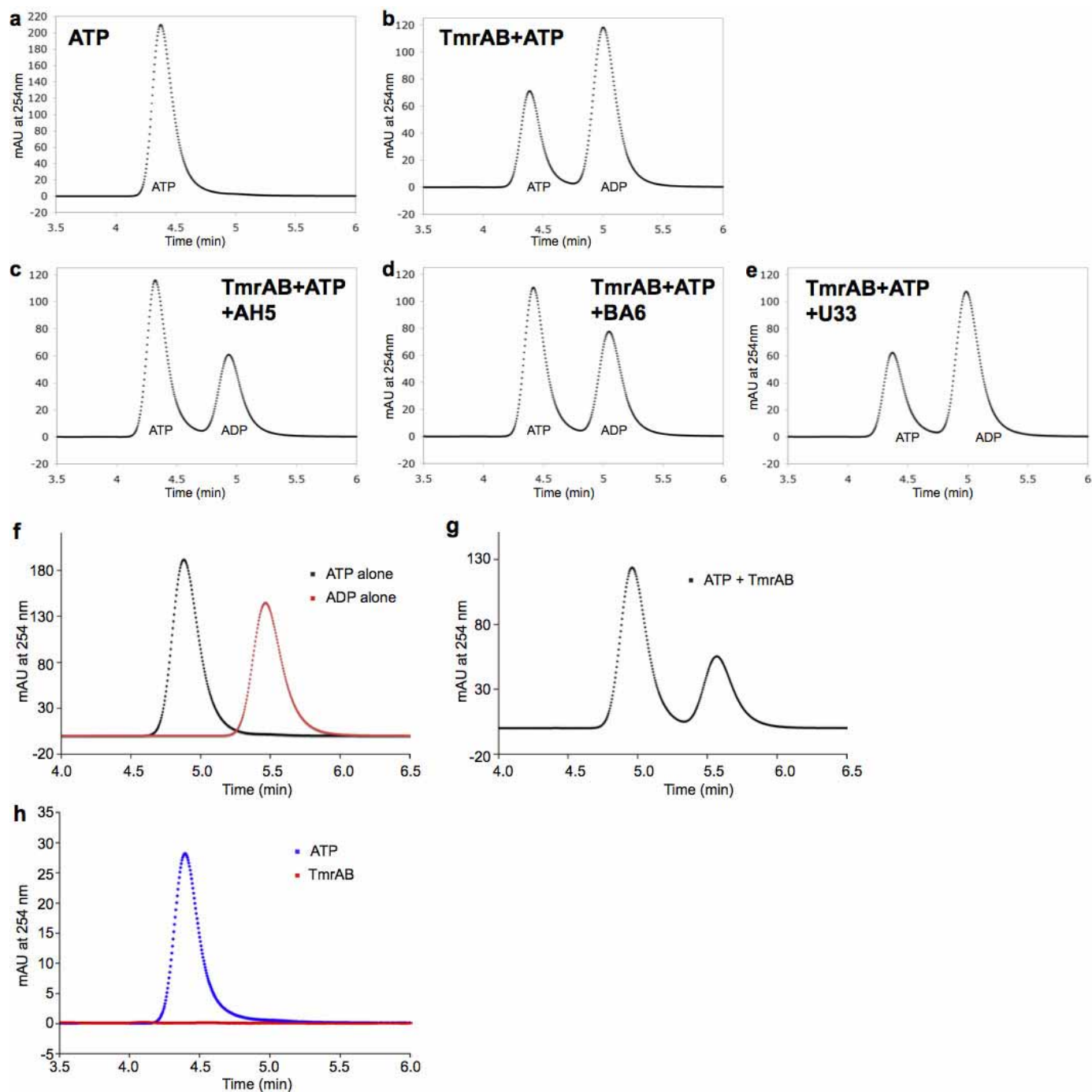
TmrAB-BA6 is coloured according to the value of local cross-correlation values of TmrAB-BA6 with TmrAB-AH5 (upper), and with TmrAB (lower). Right: density map of TmrAB is coloured according to the local cross-correlation value of TmrAB with TmrAB-AH5 (upper) and with TmrAB-BA6 (lower).



**Extended Data Figure 7 | Atomic model of TmrAB.** **a, b,** Two different views of the atomic model of TmrAB, generated by flexible fitting of the sequence homology model of TmrAB into the density map of the TmrAB-AH5 complex.

TmrA is coloured in salmon, and TmrB is coloured in blue. Intracellular loop 4 is coloured in green. **c,** Two subunits are arranged with a pseudo-two-fold symmetry.

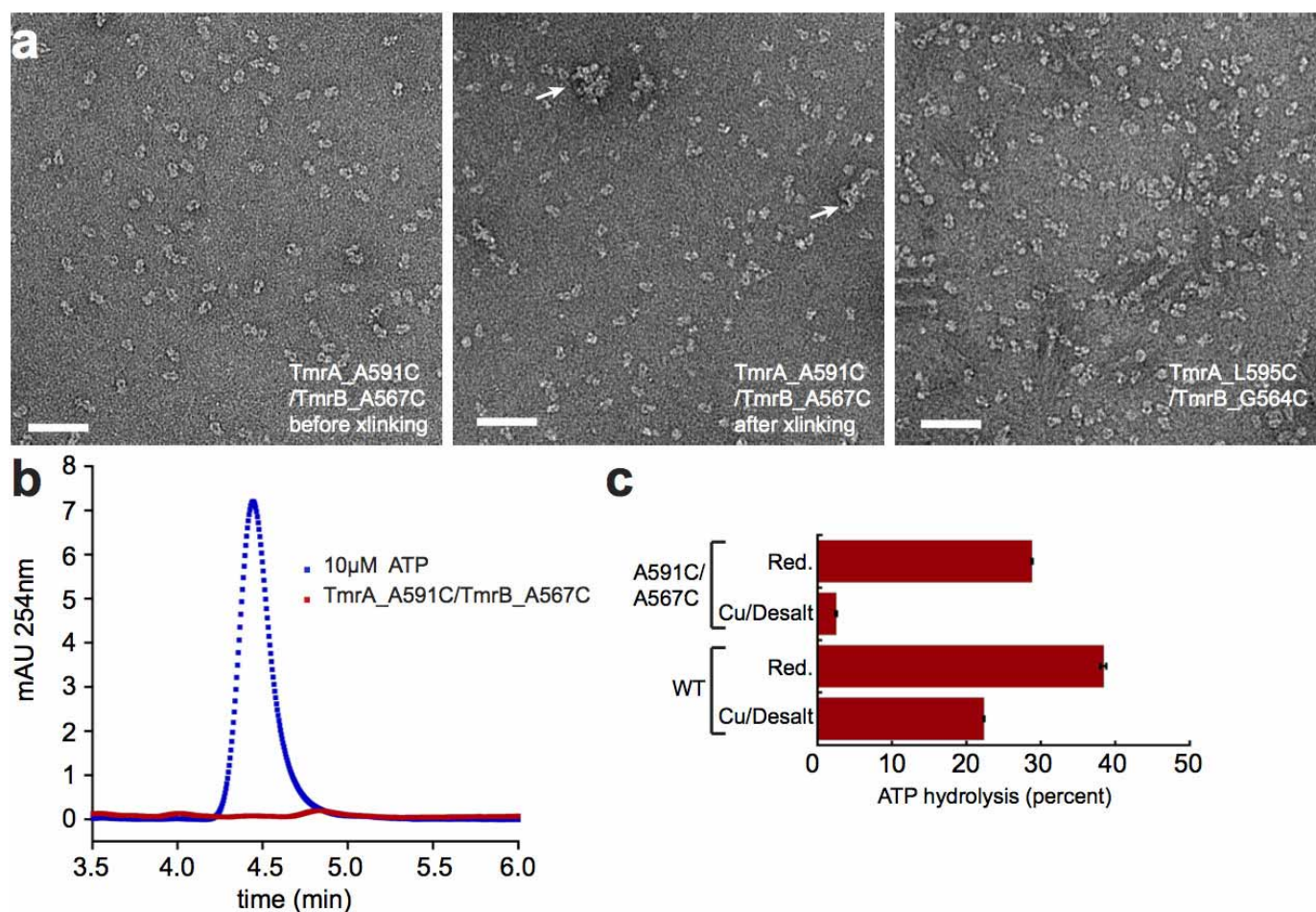




**Extended Data Figure 8 | AH5 and BA6 inhibit the ATPase activity of the TmrAB.** **a**, ATP standard for **b–e**. **b–e**, ATP hydrolysis assay at 37 °C. Reactions were performed at 37 °C for 20 min with 6.25  $\mu$ M of TmrAB, 250  $\mu$ M ATP and 2 mM  $\text{MgCl}_2$  in the presence of 25  $\mu$ M of AH5 (**c**), BA6 (**d**) or a negative control Fab, U33 (**e**). ATP hydrolysis by TmrAB was reduced in the presence of AH5 or BA6 compared with the equivalent reaction in the absence of Fabs (**b**). ATP hydrolysis was not affected by the presence of U33, which does not bind to TmrAB (**e**). **f**, ATP and ADP standards (250  $\mu$ M each)

for **g**. Two peaks were resolved corresponding to ATP and ADP (black and red curves respectively). **g**, ATP hydrolysis by TmrAB was performed with 70 nM of TmrAB, 250  $\mu$ M ATP and 2 mM  $\text{MgCl}_2$  at 60 °C for 30 min. **h**, Identification of the TmrAB nucleotide-binding state. ATP was not detected from the protein-extracted aqueous phase (red curve). ATP at an equivalent concentration (blue curve) is shown as a control to demonstrate sufficient sensitivity for nucleotide detection.





**Extended Data Figure 9 | Cysteine cross-linking validating the interaction between the C-terminal helices of TmrAB in the nucleotide-free state.**

**a**, Three samples (marked with an asterisk in Fig. 3d) were visualized by negative-stain electron microscopy, showing that TmrAB with the double cysteine mutation has the native dimeric shape of TmrAB. TmrAB contains an exposed native cysteine residue (TmrA-C416) that could not be removed. It causes some inter-dimer cross-linking (marked by arrows) under the

oxidative condition. Scale bar, 50 nm. **b**, Analytical HPLC demonstrating that purified TmrAB containing the A591C/A567C mutation is nucleotide free.

**c**, ATP hydrolysis assay indicating that disulphide cross-linking inhibits the ATPase activity of TmrAB containing the double cysteine mutation. Assays were performed in triplicate for 1 h at 60 °C with 70 nM reduced or oxidized TmrAB, 250 μM ATP, and 2 mM MgCl<sub>2</sub> before analysis by analytical HPLC.

**Extended Data Table 1 | Summary of TmrAB structure determination by single-particle cryo-EM**

Protein sample	TmrAB-AH5	TmrAB-AH5	TmrAB-BA6	TmrAB
Electron microscope	TF20	TF30 Polara	TF20	TF20
Pixel size (Å)	3.74Å	1.96Å	3.74Å	3.74Å
Number of micrograph	347	639	301	594
# of particles picked	27,000	131,000	41,000	46,000
Final # of particles	27,000	102,000	30,000	36,000
Resolution (Å)	10.6Å	8.2Å	9.4Å	10.0Å
AccuracyRotations*	6.579°	7.306°	7.094°	9.5°
AccuracyTranslation*	1.085 pixel	2.018 pixel	0.994 pixel	1.195 pixel
AccuracyTranslation*	4.06Å	3.96Å	3.72Å	4.47Å
B-factor	N/A	-1195Å <sup>2</sup>	-1415Å <sup>2</sup>	-1851Å <sup>2</sup>

We determined three three-dimensional reconstructions of TmrAB complexed with a Fab, AH5 or BA6, and a three-dimensional reconstruction of TmrAB alone, using two different microscopes, TF30 Polara and TF20. TF30 Polara was equipped with a K2 Summit camera, and TF20 with an 8k × 8k phosphor scintillator based CMOS camera. All three-dimensional reconstructions were determined using RELION. Parameters of accuracy (marked with an asterisk) were obtained from RELION. The rotational accuracy of refinement of TmrAB in complex with Fabs (~6.5–7.3°) is significantly better than the TmrAB alone (9.5°), suggesting that the Fabs improved the accuracy of angular refinement.

# CAREERS

**RESPONSIBILITY** Royal Society urges students and advisers to explore all job options **p.403**

**TRANSLATION** Bridge-building between academia and the clinic [go.nature.com/ielkkf](http://go.nature.com/ielkkf)

**NATUREJOBS** For the latest career listings and advice [www.naturejobs.com](http://www.naturejobs.com)



CHRIS RYAN/NATURE

## WORK-LIFE BALANCE

# Lab life with kids

*Balancing research with raising children takes scheduling skills and organization.*

BY KENDALL POWELL

While she was in the middle of harvesting plates of cells at biotechnology company Genentech in South San Francisco, California, molecular oncologist Ingrid Wertz received a phone call. It was her childcare provider, telling her that her then-9-year-old daughter had a concussion after smacking heads with another child.

Wertz had been expecting to spend the next four hours processing cells as part of an experiment on cytokine signalling, but instead she found herself rushing to the childcare centre. Luckily, her work was not ruined — before she left, she managed to store it in a freezer so that she could pick up the experiment the next day.

Wertz's experience is familiar to early-career

researchers who are the parents of young children. Juggling the responsibilities of laboratory and home is difficult at best, especially because childbearing and child-rearing years tend to collide with the formative stages of a young researcher's career. The pressure to produce at the bench combined with the duties of childcare can push stress levels to breaking point.

To cope with the extra responsibilities of childcare, scientist parents must learn to partition their days and nights to accommodate work and family, and to structure experimental protocols so that they do not skimp on quality family time. For example, instead of scheduling 12-hour data-collection time points at 8 a.m. and 8 p.m., which are prime breakfast and dinner hours, they can shift them to 10 a.m. and 10 p.m., stopping back in the lab after the

children are in bed. Organizational skills are important, such as coordinating working days and nights with other people providing care. Family-friendly national or workplace policies, such as reduced teaching responsibilities and flexible hours, can ease the burden. But busy parents should also make time for themselves — whether that time is for a 20-minute jog or a date together.

## UNCERTAINTY EXPECTED

Lab work is inherently erratic: experiments can take twice as long as planned or cells may not grow, shifting schedules by a day or more. But the addition of children to a scientist's life adds unpredictability to the already unpredictable. Wertz learned the hard way that she has to plan for the unforeseen: her strategies ►



► include scheduling in lengthy buffer zones for large or crucial experiments.

She also builds in multiple back-up plans. In emergencies, such as when her son shoved an acorn up his nose at age three, the scientists in her group all lend a hand. But back-up planning also involves “knowing the non-critical times when you can stop in the middle of an experiment,” she says. Wertz works out ahead of time how to wind things down quickly without ruining data. That way, if the experiment starts to run behind schedule, it can literally be put on ice until the next day: stable DNA or proteins can be safely tucked away in a freezer.

When a dance recital or school art show waits at the other end of an experiment, Wertz's success depends on time micromanagement. “Plan every step a day ahead of time, locate all the reagents and schedule incubations during times when you are going to be called away to meetings,” she counsels.

If research has to interrupt family time, it is best to prepare the family in advance. As a clinical researcher at the University of Texas Southwestern Medical Center in Dallas, Jane Wigginton's work centres on testing interventions for patients suffering from trauma — who tend to come into the hospital late on weekend nights and during holiday celebrations. Once she was paged just before a Fourth of July fireworks display and had to leave her family.

But she and her ex-husband made sure that all six kids knew that even if an event such as Thanksgiving dinner had to be postponed, it would still happen. Looking back, she sees that the disruptions provided a good lesson for family members. “They learned that it does not have to be the exact hour or day to celebrate a family moment.”

Jens Schuster, a molecular biologist at Uppsala University in Sweden, and his wife, who works as a nurse supervisor, have also learned the art of über-organization. Recently, their 7-year-old daughter was too ill to attend school, prompting a shift in schedules. Schuster's wife stayed home with their daughter until her hospital shift started. By that point, their 17-year-old son was home from school and could watch his sister (and 12-year-old brother) for a couple of hours. Schuster swung home from the lab a bit early, picking up their 3-year-old daughter from day care on his way.

Research allows for flexible schedules, he says. “Cells don't care if you come in at midnight or at noon to take care of them.” But he constantly checks his and his wife's schedules to sidestep disasters as much as possible. Each week, on the family's shared Google calendar (an almost universal tool for researcher families), he reviews upcoming obligations, such as his work meetings, his wife's shifts and the children's dentist appointments.

He finds and resolves scheduling conflicts, then plans experiments for days that have at least four hours of uninterrupted time. By scheduling such blocks about every two weeks, he can stay on track with experiments such as coaxing stem cells into more-specialized types or analysing images on a sophisticated microscope, he says. “I wonder myself, quite often, how do I do all this?” he says.

### PRIORITY CLASH

Researcher parents concede that work–life boundaries can blur when the demands of childcare and career advancement peak simultaneously. The average age at which US researchers gain tenure is 39, when women's



Grzegorz Wicher snaps a selfie with his family.

fertility has declined sharply. Most early-career academic scientists who also want to become parents pace themselves by the biological clock and have children before they earn tenure. Those who juggle tenure committees and babies say that a shift in perspective has helped them to cope with the tension.

Rebecca Richards-Kortum, a bioengineer at Rice University in Houston, Texas, found it stressful trying to figure out the optimal timing for her first child during her pre-tenure days at another institution. She realized after a couple of years that she would regret not having a family more than not making tenure. “That was a real clarifying moment,” she says. “It helped me let go of a lot of the stress.” She went on to earn tenure and become an award-winning scientist and mother of six.

But generous institutional services also smoothed the way for her to juggle lab life and childcare obligations. Subsidized, on-site day care was key, especially when she was breastfeeding (see ‘Tenure travail’). Some of her best impromptu discussions with colleagues about grant proposals happened in the day-care centre at pick-up times.

Generous family policies have also helped Grzegorz Wicher, who is a senior postdoctoral researcher at Uppsala University. For each child, the Swedish government guarantees 13 months of leave at 80% salary, with the time off work being split between parents as desired. Some academic departments chip in another 10%. Even so, research careers can suffer from long publication gaps and breaks in momentum, says Wicher, who is also in the middle of starting up a cell-culture company.

## TENURE TRAVAIL

### *The cost of motherhood*

It is a fact. Having children comes at a higher cost to a woman's academic science career than to a man's. That finding prompted all ten campuses of the University of California in 2003 to adopt family-friendly policy reforms, such as giving mothers two semesters and fathers one semester without teaching responsibilities and automatically extending the allowed time in which to gain tenure by one year after the birth of each child.

The policies were brought in after a report ([go.nature.com/pbhjyd](http://go.nature.com/pbhjyd)) produced for the university by Mary Ann Mason and her colleagues at the University of California, Berkeley, showed that although every academic researcher is busy, women with children were working the most, devoting more than 100 hours to work, domestic and child-care responsibilities each week

(compared with fewer than 80 hours per week for faculty members without children).

Mason's research showed that having children within five years of obtaining a PhD lowers the chances that women will enter a tenure-track position and earn tenure. Women who have children more than five years after finishing their PhD do as well as women without children, but there are far fewer of these “late babies”, says Mason, who is a lawyer and professor at Berkeley's graduate school.

For her book *Mothers on the Fast Track*, Mason interviewed successful mothers from many fields and found that the most important contributing factor to their success was a partner who felt that their spouse's career was as important as their own. “My advice in the book,” she says, “is don't marry a jerk.” **K.P.**



To maximize efficiency, he and his wife, who is also a senior postdoc at Uppsala, planned their research around their respective portions of parental leave after the birth of their daughter, now aged 8 months. They each arranged to finish laboratory projects before taking leave, and planned to use nap times and evenings at home to work on data analysis, manuscripts and grant proposals.

In their house, mornings at the breakfast table are sacrosanct family time; so are the hours after their 6-year-old daughter's school day and until the children's bedtime. The couple typically works side-by-side in their home office for three to four hours after that. "It is a little bit sad, but better than not seeing each other at all," Wicher says.

## WORKING LATE

In the absence of parental-leave or child-care policies, scientist parents turn to other strategies to accommodate lab obligations and family time. Many with young children split up their days and nights, returning to the lab during the late evening and working remotely when possible. Anthony Barry, an associate research fellow at Pfizer Biotherapeutics in Andover, Massachusetts, takes his laptop home every evening.

"I get incredibly frustrated if I get home so late that I'm not getting to see my kids," says Barry, whose sons are aged 7 and 10. Dividing his duties into work that must be done at Pfizer versus what can be done from home helps him to complete 8–10-hour workdays without missing prime family time. "Although people may say it's horrible to have to take work home with you, I've found that to be the most enabling," says Barry.

Others see the evening hours as the perfect time to head back to the lab. Amy Pandya-Jones, a postdoctoral researcher in RNA biology at the University of California, Los Angeles, splits her days to get quality time with her 5-year-old and 2-year-old. She goes to the lab early in the morning and comes home in the early afternoon. About three nights per week, after her husband gets home from work at around 7 p.m., she returns to the lab, working for another four hours.

She is careful to waste not a second, and estimates that she squeezes what would normally be a full 8–10-hour workload into about 6–8 hours. "You cannot underestimate the planning," she says: she slots in time on the weekly calendar even for a trip to the supermarket.

Parents who manage to carve out minutes for themselves and their partners relieve some of the stress. One practice that helped Wigginton to stay sane was stealing an hour or two for herself, sometimes for a manicure or a pedicure. "I needed just a moment away, with nobody sitting in my lap and no pager going off."

Wicher and Pandya-Jones both reserve

one night a week for dates with their spouses — even if it is just a dinner of tacos and beer. Wicher and his wife also take it in turns to go running on alternate days. "It helps to wash the brain," he says. Jaelyn Eberle, a palaeontologist at the University of Colorado, Boulder, recently finished a series of exercise sessions that started at 5:15 a.m. in the mornings. "I realize that if I don't get some me time — exercise or pottery lessons — then I'm not as creative at work," she says.

On Sunday evenings when the kids are asleep, Wertz and her husband serve themselves ice cream and sit down to look over their family calendar, plan, organize and talk. "We make a fun time of it — we get different Ben and Jerry's flavours and sample the new ones."

Some researchers hire house cleaners and sitters, ask neighbours to drive kids to activities or order groceries online to allocate their limited hours at home to family time rather than chores. It pays to ask for help from friends, relatives and even employers — especially for single-parent scientists who have less support at home. That could mean asking grandparents to babysit for a weekend so that the researcher can finish up a grant application or asking a boss for a month's notice before scheduling a business trip.

But it is not all about nappy duty, day care and drudge. Researchers see benefits

**"Cells don't care if you come in at midnight or at noon to take care of them."**

for themselves and for their children from their work. When the children were older, Wigginton took one or two of them (and

her mother) on conference trips to Paris or Hawaii. Wertz enjoys "watching the joy and fascination, through the eyes of a child, of ice melting and water pouring" as her kids play in her lab on weekend visits.

Richards-Kortum believes that blending research and parenting strengthens both endeavours. Her experience as a mother has helped to shape her research agenda on life-saving technologies for premature newborns. And her work in Malawi influenced her decision to adopt her two youngest daughters from Ethiopia.

Both she and Richards-Kortum have evidence that the hours devoted to research did not leave their children feeling resentful towards their scientific research careers — all of their university-aged children are following in their mothers' footsteps, studying engineering, bioengineering or medicine. ■

**Kendall Powell** is a freelance writer in Lafayette, Colorado.

## CAREER PROGRESSION

### Consider all options

Research institutions should provide broader career guidance to their PhD students, and students should proactively assess their skills and options, according to a report by the Royal Society, an influential group of scientists based in London.

"Students must not be regarded as mere 'bench monkeys', but nor should they themselves be passive in seeking out what they need," wrote Athene Donald, who chaired the group that put together the report, in an accompanying opinion piece.

The Royal Society is certainly not the first science organization to highlight the grim chances of newly minted graduates and postdocs finding faculty positions in scientific research and to call for universities to provide better career preparation (see *Nature* **516**, 7–8; 2014). A report from the US National Academies in Washington DC, for example, says that postdoctoral positions, often seen as the default step after a PhD programme, do not always help researchers to advance their careers, and that research institutions should inform PhD students that other types of work experience may be more beneficial (see [go.nature.com/cxli6t](http://go.nature.com/cxli6t)).

Donald, a theoretical physicist at the University of Cambridge, says that the report, entitled *Doctoral Students' Career Expectations: Principles and Responsibilities*, aims to raise awareness of viable career options among students and their supervisors, and to bolster efforts by university career-guidance offices. Improving career awareness may require students and schools to arrange mentorships beyond a trainee's lab, department or institution. And PhD advisers should not imply that a future in academia is the only desirable career path (see [go.nature.com/h9872d](http://go.nature.com/h9872d)).

Lack of information is a serious issue, but merely highlighting careers beyond academia may not do much to help people to find optimal positions, says Sally Hancock, a higher-education researcher at the University of York, UK, who studied science PhD students at Imperial College London. Those who had been exposed only to academic research were likely to be purists who saw their programme as "a zero-sum game in which the objective is to achieve an academic position". Those who had already worked outside academia were more likely to view non-academic options favourably, be proactive about exploring other choices and feel less stigma about pursuing them. PhD programmes might serve students best by incorporating work outside the university, she says: "Experience is imperative."

# THE LEFT HANDS OF LOVERS

*Strong connection.*

BY EMILY ECKART

The deactivated robots were corralled upright in cages. I navigated the rows under the red light, searching for the only face that mattered. She was in the back. Her shapely features stood out; the other robots had been designed ugly on purpose.

"Rosalind," I whispered, foolishly expecting a response. But there was only my breath and the soft whine of the lights: no more of her pleasant hum, reassuring as a heartbeat. She was fully deactivated, her memory chip wiped clean.

Even in deactivation, her face held that mysterious semi-smile. I realized it was her default expression — not, as I had believed, a smile just for me.

I met Rosalind the week after the terrorists fled, when the city still gleamed with blood and broken glass. The commander introduced the robots to our unit. They resembled both sexes and all races, but were uniformly ugly — eyes too close together, odd-shaped noses, warts.

"You'll use these robots for inspecting buildings abandoned by the terrorists," he said. "Their exteroceptive sensors can detect the vapours of eight different explosives, as well as chemical weapons like nerve agents. That way we don't risk human lives." He assigned each soldier a robot. When he got to me, there was none left.

He took me aside. "There's one robot left from the previous generation. Not an ideal model, but they're too expensive to waste. We thought you could handle it."

He led me into another room. The robot was wearing jeans and a V-neck. She had skin the colour of a coffee-flavoured soy drink. Her eyes were wide and green.

"This is RL38501X. We ordered humanoid robots after studies showed soldiers worked best with machines that they related to. But this generation was too relatable. People got attached. Last year a man followed his robot inside a booby-trapped building and died."

The robot smiled at me. Her kind expression reminded me of my mother, who had been killed with the rest of my family in the bombing of '32.

➔ NATURE.COM

Follow Futures:

Twitter @NatureFutures

Facebook go.nature.com/mtoodm

"It's a machine," he said. "Think of it as it."

As I directed the SmartCar to the edge

of the city, RL38501X made conversation from her hermetically sealed compartment in the back.

"Are you from this area?" she asked.

"Nope. Transferred here to join the military back in '33."

"Do you like it?"

I shrugged. "Sometimes it's lonely."



"That is sad. No one should feel lonely."

I let her out at the checkpoint. *It*, I reminded myself. *Its* job was to investigate a suspected storage facility for Novichok agent powders. As I waited, I found myself hoping she would be safe. *Be safe, RL ...* Who could remember that serial number? RL should signify a name, I decided. Rosalind.

While other robots got blown up or contaminated by toxic chemicals, Rosalind went months without incident. We were transferred closer to the battlefield. On our first anniversary, I drove her to the hill above the new city. We looked out over the scarred buildings, the roads strewn with litter and wreckage. In the morning light, if you forgot what it meant, the glittering debris almost looked beautiful.

"Your hand is like mine," Rosalind said. I had wondered when she'd notice. My left hand was a prosthesis. I had the same thin line on my wrist as she, the same titanium joints beneath artificial skin.

"I lost my real hand. In an accident." Fleeing my besieged hometown, I had fallen in the road. My hand was crushed by an errant SmartCar.

Rosalind furrowed her eyebrows. "What is a 'real hand'? Is this not real?"

She touched my prosthesis. A static shock passed between us, sharp and bright, much more vivid than the faint sensations I was used to.

"Ow!" I felt guilty as soon as her face assumed a hurt expression. I grabbed her right hand with my left. She smiled. Quietly, we watched the sun rise higher over the city.

I felt nervous about the squat concrete building Rosalind had to inspect that week. Its broken windows gaped like missing teeth in the grin of some delinquent.

As I waited, I thought about where I'd take her that night. I had been checking her out of the robot storage room a couple of nights a week, using maintenance as an excuse. So far we had listened to a Mozart symphony and watched a Truffaut film. She hadn't yet seen the stars.

I stared out the windshield, wondering when she would return. It had been longer than usual. A call came in on my earpiece.

"Your robot has been contaminated. Return to base immediately."

"Contaminated? How?"

"A new strain of bacteria engineered by the terrorists. Extremely antibiotic resistant."

"But I can't just leave —"

"Return to base immediately. Your robot will be sent to BioLabs for analysis and disposed of safely. You'll be assigned a new one tomorrow."

The disposal facility was five miles from base, but adrenaline made it feel close. I snuck into the basement and found her crammed in a cage like trash. I reached through the bars and took her dangling hand, drawing it towards me. Her silicone skin was perfect, poreless, cold.

With my right hand, I unscrewed my left prosthesis and dropped it on the floor. I unscrewed her left hand and attached it to my stump. As I walked away, my breath came shorter. Did I feel an itching in my chest, a dizziness? Would they find my body pockmarked and ridden with buboes? I did not care. Without Rosalind, I had no one. ■

**Emily Eckart** studied music and English literature at Harvard University. Her fiction has appeared in *Potomac Review*, *Literary Orphans* and elsewhere. You can read more of her work at [www.emilyeckart.com](http://www.emilyeckart.com).

ILLUSTRATION BY JACEY